


Article

Information Geometric Approach on Most Informative Boolean Function Conjecture

Albert No 

Department of Electronical and Electrical Engineering, Hongik University, Seoul 04066, Korea; albertno@hongik.ac.kr; Tel.: +82-2-320-1649

Received: 26 July 2018; Accepted: 8 September 2018; Published: 10 September 2018

Abstract: Let X^n be a memoryless uniform Bernoulli source and Y^n be the output of it through a binary symmetric channel. Courtade and Kumar conjectured that the Boolean function $f : \{0, 1\}^n \rightarrow \{0, 1\}$ that maximizes the mutual information $I(f(X^n); Y^n)$ is a dictator function, i.e., $f(x^n) = x_i$ for some i . We propose a clustering problem, which is equivalent to the above problem where we emphasize an information geometry aspect of the equivalent problem. Moreover, we define a normalized geometric mean of measures and interesting properties of it. We also show that the conjecture is true when the arithmetic and geometric mean coincide in a specific set of measures.

Keywords: Boolean function; Bregman divergence; clustering; geometric mean; Jensen–Shannon divergence

1. Introduction

Let X^n be an independent and identically distributed (i.i.d.) uniform Bernoulli source and Y^n be an output of it through a memoryless binary symmetric channel with crossover probability $p < 1/2$. Recently, Courtade and Kumar conjectured that the most informative Boolean function is a dictator function.

Conjecture 1 ([1]). *For any Boolean function $f : \{0, 1\}^n \rightarrow \{0, 1\}$, we have:*

$$I(f(X^n); Y^n) \leq 1 - h_2(p) \quad (1)$$

where the maximum is achieved by a dictator function, i.e., $f(x^n) = x_i$ for some $1 \leq i \leq n$. Note that $h_2(p) = -p \log p - (1 - p) \log(1 - p)$ is the binary entropy function.

Although there has been some progress in this line of work [2,3], this simple conjecture still remains open. There are also a number of variations of this conjecture. Weinberger and Shayevitz [4] considered the optimal Boolean function under quadratic loss. Huleihel and Ordentlich [5] considered the complementary case and showed that $I(f(X^n); Y^n) \leq (n - 1)(1 - h_2(p))$ for all $f : \{0, 1\}^n \rightarrow \{0, 1\}^{n-1}$. Nazer et al. focused on information distilling quantizers [6], which can be seen as a generalized version of the above problem.

Many of them are based on the Fourier analysis technique including the original paper [1]. In this paper, we suggest an alternative approach, namely the information geometric approach. The mutual information can naturally be expressed with Kullback–Leibler (KL) divergences. Thus, it can be shown that the maximizing mutual information is equivalent to clustering probability measures under KL divergence.

In the equivalent clustering problem, the center of the cluster is an arithmetic mean of measures. We also provide the role of the geometric mean of measures (with appropriate normalization) in this

setting. To the best of our knowledge, the geometric mean of measures has received less attention in the literature. We propose an equivalent formulation of the conjecture using the geometric mean of measures. Note that the geometric mean also allows us to connect Conjecture 1 to the other well-known clustering problem.

The rest of the paper is organized as follows. In Section 2, we briefly review the Jensen–Shannon divergence and \mathcal{I} -compressedness. In Section 3, we provide an equivalent clustering problem of probability measures. We introduce the geometric mean of measures in Section 4. We conclude this paper in Section 5.

Notations

\mathcal{X} denotes the alphabet set of random variable X , and $\mathcal{M}(\mathcal{X})$ denotes the set of measures on \mathcal{X} . X^n denotes a random vector (X_1, X_2, \dots, X_n) , while x^n denotes a specific realization of it. If it is clear from the context, $P_{Y|X}$ denotes a conditional distribution of Y given $X = x$, i.e., $P_{Y|X}(y) = P_Y(y|x)$. Similarly, $P_{Y^n|X^n}$ denotes a conditional distribution of Y^n given $X^n = x^n$, i.e., $P_{Y^n|X^n}(y^n) = P_{Y^n|X^n}(y^n|x^n)$. Let $\Omega = \{0, 1\}^n$ be the set of all binary sequences of length n . For $A \subseteq \Omega$, the shifted version of A is denoted by $A \oplus x^n = \{\tilde{x}^n \oplus x^n : \tilde{x}^n \in A\}$ where \oplus is an element-wise XOR operator. The arithmetic mean of measures in the set $\{P_{Y^n|X^n} : x^n \in A\}$ is denoted by μ_A . For $1 \leq i \leq n$, let A_{i0} be the set of elements in A that satisfy $x_i = 0$, i.e., $A_{i0} = \{x^n \in A : x_i = 0\}$, and $\Omega_{i0} = \{x^n \in \Omega : x_i = 0\}$. A_{i1} is defined in a similar manner. A length n binary vector $x^{n-1}0$ denotes a vector x^n with $x_n = 0$.

2. Preliminaries

2.1. Jensen–Shannon Divergence

For $\alpha_1, \alpha_2 \geq 0$ such that $\alpha_1 + \alpha_2 = 1$, the Jensen–Shannon (JS) divergence of two measures P_1 and P_2 is defined as:

$$\text{JSD}_\alpha(P_1, P_2) = H(\alpha_1 P_1 + \alpha_2 P_2) - \alpha_1 H(P_1) - \alpha_2 H(P_2). \quad (2)$$

It is not hard to show that the following definition is equivalent.

$$\text{JSD}_\alpha(P_1, P_2) = \alpha_1 D(P_1 \| \alpha_1 P_1 + \alpha_2 P_2) + \alpha_2 D(P_2 \| \alpha_1 P_1 + \alpha_2 P_2). \quad (3)$$

Lin proposed a generalized JS divergence [7]:

$$\text{JSD}_\alpha(P_1, P_2, \dots, P_n) = H\left(\sum_{i=1}^n \alpha_i P_i\right) - \sum_{i=1}^n \alpha_i H(P_i) \quad (4)$$

where $\alpha = (\alpha_1, \dots, \alpha_n)$ is a weight vector such that $\sum_{i=1}^n \alpha_i = 1$. Similar to Equation (3), it has an equivalent definition:

$$\text{JSD}_\alpha(P_1, P_2, \dots, P_n) = \sum_{i=1}^n \alpha_i D(P_i \| \bar{P}) \quad (5)$$

where $\bar{P} = \sum_{i=1}^n \alpha_i P_i$. Topsøe [8] pointed out an interesting property, the so-called compensation identity. It states that for any distribution Q ,

$$\sum_{i=1}^n \alpha_i D(P_i \| Q) = \sum_{i=1}^n \alpha_i D(P_i \| \bar{P}) + D(\bar{P} \| Q) \quad (6)$$

$$= \text{JSD}_\alpha(P_1, P_2, \dots, P_n) + D(\bar{P} \| Q). \quad (7)$$

Throughout the paper, we often use Equation (6) directly without the notion of JSD

Remark 1. The generalized JS divergence is the mutual information between X and the mixture distribution. Let Z be a random variable that takes the value from $\{1, 2, \dots, n\}$ where $P_Z(i) = \alpha_i$ and $P_{X|Z}(x|i) = P_i(x)$. Then, it is not hard to show that:

$$\text{JSD}_\alpha(P_1, P_2, \dots, P_n) = I(X; Z) \quad (8)$$

However, we introduced generalized JS divergence to emphasize the information geometric perspective of our problem.

2.2. \mathcal{I} -Compressed

Let A be the subset of Ω and $\mathcal{I} = \{i_1, i_2, \dots, i_k\} \subset \{1, 2, \dots, n\}$ be the set of indexes. For $x^n \in \Omega$, the \mathcal{I} -section of A is defined as:

$$A_{\mathcal{I}}(x^n) = \left\{ z^k : y^n \in A, y_i = \begin{cases} z_j & \text{if } i = i_j \in \mathcal{I} \\ x_i & \text{otherwise} \end{cases} \right\}. \quad (9)$$

The set A is called \mathcal{I} -compressed if $A_{\mathcal{I}}(x^n)$ is an initial segment of lexicographical ordering for all x^n . For example, if A is \mathcal{I} -compressed for some $|\mathcal{I}| = 2$, then $A_{\mathcal{I}}(x^n)$ should be one of:

$$\{00\}, \{00, 01\}, \{00, 01, 10\}, \{00, 01, 10, 11\}. \quad (10)$$

It simply says that if $x^{n-2}10 \in A$, then $x^{n-2}00, x^{n-2}01 \in A$.

Courtade and Kumar showed that it is enough to consider the \mathcal{I} -compressed sets.

Theorem 1 ([1]). Let \mathcal{S}_n be the set of functions $f : \Omega \rightarrow \{0, 1\}$ for which $f^{-1}(0)$ is \mathcal{I} -compressed for all \mathcal{I} with $|\mathcal{I}| \leq 2$. In maximizing $I(f(X^n); Y^n)$, it is sufficient to consider functions $f \in \mathcal{S}_n$.

In this paper, we often restrict our attention to functions in the set \mathcal{S}_n .

3. Approach via Clustering

In this section, we provide an interesting approach toward Conjecture 1 via clustering. More precisely, we formulate an equivalent clustering problem.

3.1. Equivalence to Clustering

The following theorem implies the relation between the original conjecture and the clustering problem.

Theorem 2. Let $f : \mathcal{X} \rightarrow \mathcal{U}$ and $U = f(X)$ be an induced random variable. Then,

$$I(f(X); Y) = I(X; Y) - \sum_x P_X(x) D(P_{Y|x} \| P_{Y|U}(\cdot | f(x))). \quad (11)$$

The proof of the theorem is provided in Appendix A. Note that:

$$P_{Y|U}(y|u) = \frac{P_{U|Y}(u|y)P_Y(y)}{P_U(u)} \quad (12)$$

$$= \frac{P_Y(y)}{P_U(u)} \sum_{x \in f^{-1}(u)} P_{X|Y}(x|y) \quad (13)$$

$$= \sum_{x \in f^{-1}(u)} \frac{P_X(x)}{P_U(u)} \cdot P_{Y|X}(y|x). \quad (14)$$

which is a weighted mean of $P_{Y|X}(y|x)$ for $x \in f^{-1}(u)$. The $D(P_{Y|x} \| P_{Y|U}(\cdot | f(x)))$ is a distance from each element to the cluster center. This implies that maximizing $I(f(X); Y)$ is equivalent to clustering $\{P_{Y^n|x^n}\}$ under KL divergences. Since KL divergence is a Bregman divergence, all clusters are separated by a hyperplane [9].

In this paper, we focus on $\mathcal{U} = \{0, 1\}$ where X^n is i.i.d. Bern(1/2).

Corollary 1. Let $f : \Omega \rightarrow \{0, 1\}$ and $U = f(X^n)$ be a binary random variable.

$$I(f(X^n); Y^n) = n - nh_2(p) - \frac{1}{2^n} \sum_{x^n} D(P_{Y^n|x^n} \| P_{Y^n|U}(\cdot | f(x^n))). \quad (15)$$

The equivalent clustering problem is minimizing:

$$\sum_{x^n} D(P_{Y^n|x^n} \| P_{Y^n|U}(\cdot | f(x^n))). \quad (16)$$

Let $A = \{x^n \in \Omega : f(x^n) = 0\}$, then we can simplify $P_{Y^n|U}$ further.

$$P_{Y^n|U}(y^n|0) = \frac{1}{|A|} \sum_{x^n \in A} P_{Y^n|X^n}(y^n|x^n) \quad (17)$$

$$\stackrel{\Delta}{=} \mu_A(y^n). \quad (18)$$

The cluster center μ_A is an arithmetic mean of measures in the set $\{P_{Y^n|x^n} : x^n \in A\}$. Then, we have:

$$\sum_{x^n} D(P_{Y^n|x^n} \| P_{Y^n|U}(\cdot | f(x^n))) = \sum_{x^n \in A} D(P_{Y^n|x^n} \| \mu_A) + \sum_{x^n \in A^c} D(P_{Y^n|x^n} \| \mu_{A^c}). \quad (19)$$

For simplicity, let:

$$\mathcal{D}(A) \stackrel{\Delta}{=} \sum_{x^n \in A} D(P_{Y^n|x^n} \| \mu_A) \quad (20)$$

which is the sum of distances from each element in A to the cluster center. In short, finding the most informative Boolean function f is equivalent to finding the set $A \subseteq \Omega$ that minimizes $\mathcal{D}(A) + \mathcal{D}(A^c)$.

Remark 2. Conjecture 1 implies that $A = \Omega_{i0} = \{x^n : x_i = 0\}$ minimizes (19). Furthermore, Theorem 1 implies that it is enough to consider A such that $A_{i1} \subseteq A_{i0}$ for all i .

For any $Q_{Y^n} \in \mathcal{M}(\Omega)$, Equation (6) implies that:

$$\sum_{x \in A} D(P_{Y^n|x^n} \| Q_{Y^n}) = \mathcal{D}(A) + |A|D(\mu_A \| Q_{Y^n}) \quad (21)$$

$$\sum_{x \in A^c} D(P_{Y^n|x^n} \| Q_{Y^n}) = \mathcal{D}(A^c) + |A^c|D(\mu_{A^c} \| Q_{Y^n}). \quad (22)$$

Thus, we have:

$$\sum_{x \in \Omega} D(P_{Y^n|x^n} \| Q_{Y^n}) = \mathcal{D}(A) + \mathcal{D}(A^c) + |A|D(\mu_A \| Q_{Y^n}) + |A^c|D(\mu_{A^c} \| Q_{Y^n}). \quad (23)$$

Note that $\sum_{x \in \Omega} D(P_{Y^n|x^n} \| Q_{Y^n})$ does not depend on A , and therefore, we have the following theorem.

Theorem 3. For any $Q_{Y^n} \in \mathcal{M}(\Omega)$, minimizing $\mathcal{D}(A) + \mathcal{D}(A^c)$ is equivalent to maximizing:

$$|A|D(\mu_A \| Q_{Y^n}) + |A^c|D(\mu_{A^c} \| Q_{Y^n}). \quad (24)$$

The above theorem provides an alternative problem formulation of the original conjecture.

3.2. Connection to Clustering under Hamming Distance

In this section, we consider the duality between the above clustering problem under the KL divergence and the clustering on Ω under the Hamming distance. The following theorem shows that the KL divergence on $\{P_{Y^n|x^n} : x^n \in \Omega\}$ corresponds to the Hamming distance on Ω .

Theorem 4. For all $x^n, \tilde{x}^n \in \Omega$, we have:

$$D(P_{Y^n|x^n} \| P_{Y^n|\tilde{x}^n}) = d_H(x^n, \tilde{x}^n) \cdot (1 - 2p) \log \frac{1-p}{p} \quad (25)$$

where $d_H(x^n, \tilde{x}^n)$ denotes the Hamming distance between x^n and \tilde{x}^n .

This theorem implies that the distance between two measures $P_{Y^n|x^n}$ and $P_{Y^n|\tilde{x}^n}$ is proportional to the Hamming distance between two binary vectors x^n and \tilde{x}^n . The proof of the theorem is provided in Appendix B. Note that the KL divergence $D(\cdot \| \cdot)$ is symmetric on $\{P_{Y^n|x^n} : x^n \in \Omega\}$.

In the above duality, we have a mapping between $\{P_{Y^n|x^n} : x^n \in \Omega\}$ and $\{0, 1\}^n$; more precisely, $P_{Y^n|x^n} \leftrightarrow x^n$. This mapping naturally suggests an equivalent clustering problem of n -dimensional binary vectors. However, the cluster center μ_A is not an element of $\{P_{Y^n|x^n} : x^n \in \Omega\}$ in general. In order to formulate an equivalent clustering problem, we need to answer the question “Which n dimensional vector corresponds to μ_A ?”. A naive approach is to extend the set of binary vectors to $[0, 1]^n$ under ℓ^2 distance instead of the Hamming distance. In such a case, the goal is to map μ_A to the arithmetic mean of binary vectors in the set A . If this is true, we can further simplify the problem into

the problem of clustering a hypercube in \mathbb{R}^n . However, the following example shows that this naive extension is not valid.

Example 1. Let $n = 2$, $A = \{00, 11\}$ and $B = \{01, 10\}$, then the arithmetic mean of binary vectors of A and that of B are the same. However, μ_A is not equal to μ_B .

Furthermore, the set Ω_{i0} is not the optimum choice when clustering the hypercube under ℓ^2 . Instead, we need to consider the set of measures directly. The following theorem provides a bit of geometric structure among measures.

Theorem 5. For all $x^n, \tilde{x}^n \in \Omega$ and $Q_{Y^n} \in \text{conv}(\{P_{Y^n|x^n} | x^n \in \Omega\})$,

$$D(P_{Y^n|X^n=x^n} \| Q_{Y^n}) - D(P_{Y^n|X^n=\tilde{x}^n} \| Q_{Y^n}) \leq k \cdot ((1-p)^k - p^k) \log \frac{1-p}{p} \quad (26)$$

where $k = d_H(x^n, \tilde{x}^n)$.

The proof of the theorem is provided in Appendix C. Since $(1-p)^k - p^k \leq 1 - 2p$ for all $k \geq 1$, Theorem 5 immediately implies the following corollary.

Corollary 2. For all $x^n, \tilde{x}^n \in \Omega$ and $Q_{Y^n} \in \text{conv}(\{P_{Y^n|x^n} | x^n \in \Omega\})$,

$$\left| D(P_{Y^n|X^n=x^n} \| Q_{Y^n}) - D(P_{Y^n|X^n=\tilde{x}^n} \| Q_{Y^n}) \right| \leq D(P_{Y^n|X^n=x^n} \| P_{Y^n|X^n=\tilde{x}^n}) \quad (27)$$

where $\text{conv}(A)$ is a convex hull of measures in the set A .

This is a triangle inequality that can be useful when we consider the clustering problem of measures.

4. Geometric Mean of Measures

In the previous section, we formulate the clustering problem that is equivalent to the original maximizing mutual information problem. In this section, we provide another approach using a geometric mean of measures. We define the geometric mean of measures formally and derive a nontrivial conjecture, which is equivalent to Conjecture 1.

4.1. Definition of the Geometric Mean of Measures

For measures $P_1, P_2, \dots, P_n \in \mathcal{M}(\mathcal{X})$ and weights $\alpha_i \geq 0$ such that $\sum_{i=1}^n \alpha_i = 1$, we considered the sum of KL divergences in (6):

$$\sum_{i=1}^n \alpha_i D(P_i \| Q). \quad (28)$$

We also observed that (28) is minimized when Q is an arithmetic mean of measures.

Since the KL divergence is asymmetric, it is natural to consider the sum of another direction of KL divergences.

$$\sum_{i=1}^n \alpha_i D(Q \| P_i) = \sum_{i=1}^n \alpha_i \sum_{x \in \mathcal{X}} Q(x) \log \frac{Q(x)}{P_i(x)} \quad (29)$$

$$= \sum_{x \in \mathcal{X}} \sum_{i=1}^n \alpha_i Q(x) \log \frac{Q(x)}{P_i(x)} \quad (30)$$

$$= \sum_{x \in \mathcal{X}} Q(x) \log \frac{Q(x)}{\prod_{i=1}^n (P_i(x))^{\alpha_i}}. \quad (31)$$

Compared to the arithmetic mean that minimizes (28), $\prod_{i=1}^n (P_i(x))^{\alpha_i}$ can be considered as a geometric mean of measures. However, $\prod_{i=1}^n (P_i(x))^{\alpha_i}$ is not a measure in general, and normalization is required. With a normalizing constant s , we can define the geometric mean of measures by:

$$\bar{P}_G(x) = \frac{1}{s} \prod_{i=1}^n (P_i(x))^{\alpha_i} \quad (32)$$

where s is a constant, so that $\sum_{x \in \mathcal{X}} \bar{P}_G(x) = 1$, i.e.,

$$s = \sum_x \prod_{i=1}^n (P_i(x))^{\alpha_i}. \quad (33)$$

Then, we have:

$$\sum_{i=1}^n \alpha_i D(Q \| P_i) = D(Q \| \bar{P}_G) + \log \frac{1}{s} \quad (34)$$

which can be minimized when $Q = \bar{P}_G$. Thus, for all Q ,

$$\sum_{i=1}^n \alpha_i D(Q \| P_i) \geq \sum_{i=1}^n \alpha_i D(\bar{P}_G \| P_i) \quad (35)$$

$$= \log \frac{1}{s}. \quad (36)$$

The above result provides a geometric compensation identity.

$$\sum_{i=1}^n \alpha_i D(Q \| P_i) = D(Q \| \bar{P}_G) + \sum_{i=1}^n \alpha_i D(\bar{P}_G \| P_i). \quad (37)$$

This also implies that $\log \frac{1}{s} \geq 0$.

Remark 3. If $n = 2$, s is called the α -Chernoff coefficient, and it is called the Bhattacharyya coefficient when $\alpha = 1/2$. The summation $\log \frac{1}{s} = \sum_{i=1}^2 \alpha_i D(\bar{P}_G \| P_i)$ is known as α -Chernoff divergence. For more details, please see [10,11] and the references therein.

Under this definition, we can find the geometric mean of measures in the set $\{P_{Y^n|\tilde{x}^n} : \tilde{x}^n \in B\}$ with uniform weights $\frac{1}{|B|}$ by:

$$\gamma_B(y^n) = \frac{1}{s_B} \left(\prod_{\tilde{x}^n \in B} P_{Y^n|X^n}(y^n|\tilde{x}^n) \right)^{1/|B|} \quad (38)$$

where:

$$s_B = \sum_{y^n} \left(\prod_{\tilde{x}^n \in B} P_{Y^n|X^n}(y^n|\tilde{x}^n) \right)^{1/|B|}. \quad (39)$$

Remark 4. The original conjecture is that the Boolean function f such that $f^{-1}(0) = \Omega_{i0} = \{x^n : x_i = 0\}$ maximizes the mutual information $I(f(X^n); Y^n)$. The geometric mean of measures in the set $\{P_{Y^n|X^n} : x^n \in \Omega_{i0}\}$ satisfies the following property.

$$\gamma_{\Omega_{i0}} = \mu_{\Omega_{i0}} \quad (40)$$

$$s_{\Omega_{i0}} = 2^{n-1} (p(1-p))^{(n-1)/2}. \quad (41)$$

Note that the geometric mean of measures in the set $\{P_{Y^n|X^n} : x^n \in \Omega\}$ satisfies:

$$\gamma_{\Omega} = \mu_{\Omega} \quad (42)$$

$$s_{\Omega} = 2^n (p(1-p))^{n/2}. \quad (43)$$

4.2. Main Results

So far, we have seen two means of measures μ_A and γ_A . It is natural to ask if they are equal. Our main theorem provides a connection to Conjecture 1.

Theorem 6. Suppose A is a nontrivial subset of $\Omega = \{0,1\}^n$ for $n > 0$ (i.e., $A \neq \emptyset, \Omega$), and A is \mathcal{I} -compressed for all $|\mathcal{I}| = 1$. Then, $A = \Omega_{i0}$ for some i if and only if $\mu_A = \gamma_A$ and $\mu_{A^c} = \gamma_{A^c}$.

The proof of the theorem is provided in Appendix D. Theorem 6 implies that the following conjecture is the equivalent to Conjecture 1.

Conjecture 2. Let $f : \Omega \rightarrow \{0,1\}$ and $A = f^{-1}(0)$ be \mathcal{I} -compressed for all $|\mathcal{I}| = 1$. Then, $I(f(X); Y)$ is maximized if and only if $\mu_A = \gamma_A$ and $\mu_{A^c} = \gamma_{A^c}$.

Remark 5. One of the main challenges of this problem is that the conjectured optimal sets are extremes, i.e., $A = \Omega_{i0}$ for some i . Our main theorem provides an alternative conjecture that seems more natural in the context of optimization.

Remark 6. It is clear that $\mu_A = \gamma_A$ holds if $|A| = 1$. Thus, both conditions $\mu_A = \gamma_A$ and $\mu_{A^c} = \gamma_{A^c}$ are needed to guarantee $A = \Omega_{i0}$ for some i .

4.3. Property of the Geometric Mean

We can derive a new identity by combining the original and geometric compensation identity together. For $A, B \subset \Omega$, let $\pi(A, B)$ be:

$$\pi(A, B) = \sum_{(x^n, \tilde{x}^n) \in A \times B} D(P_{Y^n|X^n} \| P_{Y^n|\tilde{x}^n}). \quad (44)$$

Then,

$$\pi(A, B) = \sum_{(x^n, \tilde{x}^n) \in A \times B} D(P_{Y^n|x^n} \| P_{Y^n|\tilde{x}^n}) \quad (45)$$

$$= \sum_{\tilde{x}^n \in B} \sum_{x^n \in A} D(P_{Y^n|x^n} \| P_{Y^n|\tilde{x}^n}) \quad (46)$$

$$= \sum_{\tilde{x}^n \in B} \left(\sum_{x^n \in A} D(P_{Y^n|x^n} \| \mu_A) + |A| D(\mu_A \| P_{Y|\tilde{x}^n}) \right) \quad (47)$$

$$= |B| \mathcal{D}(A) + |A| \sum_{\tilde{x}^n \in B} D(\mu_A \| P_{Y|\tilde{x}^n}) \quad (48)$$

where (47) is because of the compensation identity (6). As we discussed in Section 4.1, the second term of the right-hand side is:

$$\sum_{\tilde{x}^n \in B} D(\mu_A \| P_{Y|\tilde{x}^n}) = \sum_{\tilde{x}^n \in B} \sum_{y^n} \mu_A(y^n) \log \frac{\mu_A(y^n)}{P_{Y^n|X^n}(y^n|\tilde{x}^n)} \quad (49)$$

$$= |B| \sum_{y^n} \mu_A(y^n) \log \frac{\mu_A(y^n)}{\left(\prod_{\tilde{x}^n \in B} P_{Y^n|X^n}(y^n|\tilde{x}^n) \right)^{1/|B|}} \quad (50)$$

$$= |B| D(\mu_A \| \gamma_B) + |B| \log \frac{1}{s_B} \quad (51)$$

$$= |B| D(\mu_A \| \gamma_B) + \sum_{\tilde{x}^n \in B} D(\gamma_B \| P_{Y^n|\tilde{x}^n}). \quad (52)$$

Finally, we have:

$$\frac{1}{|A||B|} \pi(A, B) = \frac{1}{|A|} \mathcal{D}(A) + D(\mu_A \| \gamma_B) + \log \frac{1}{s_B} \quad (53)$$

$$= \frac{1}{|A|} \sum_{x^n \in A} D(P_{Y^n|x^n} \| \mu_A) + D(\mu_A \| \gamma_B) + \frac{1}{|B|} \sum_{\tilde{x}^n \in B} D(\gamma_B \| P_{Y^n|\tilde{x}^n}). \quad (54)$$

More interestingly, we can apply original and geometric compensation identities:

$$\frac{1}{|A||B|} \pi(A, B) = \frac{1}{|A|} \sum_{x^n \in A} D(P_{Y^n|x^n} \| \mu_A) + \frac{1}{|B|} \sum_{\tilde{x}^n \in B} D(\mu_A \| P_{Y^n|\tilde{x}^n}) \quad (55)$$

$$= \frac{1}{|A|} \sum_{x^n \in A} D(P_{Y^n|x^n} \| \gamma_B) + \frac{1}{|B|} \sum_{\tilde{x}^n \in B} D(\gamma_B \| P_{Y^n|\tilde{x}^n}). \quad (56)$$

From Theorem 4, we have $\pi(A, B) = \pi(B, A)$, and therefore, we can switch A and B .

$$\frac{1}{|A||B|} \pi(A, B) = \frac{1}{|B|} \sum_{x^n \in B} D(P_{Y^n|x^n} \| \mu_B) + D(\mu_B \| \gamma_A) + \frac{1}{|A|} \sum_{\tilde{x}^n \in A} D(\gamma_A \| P_{Y^n|\tilde{x}^n}) \quad (57)$$

$$= \frac{1}{|B|} \sum_{x^n \in B} D(P_{Y^n|x^n} \| \mu_B) + \frac{1}{|A|} \sum_{\tilde{x}^n \in A} D(\mu_B \| P_{Y^n|\tilde{x}^n}) \quad (58)$$

$$= \frac{1}{|B|} \sum_{x^n \in B} D(P_{Y^n|x^n} \| \gamma_A) + \frac{1}{|A|} \sum_{\tilde{x}^n \in A} D(\gamma_A \| P_{Y^n|\tilde{x}^n}). \quad (59)$$

If we let $A = B$, we have:

$$\frac{1}{|A|^2} \pi(A, A) = \frac{1}{|A|} \sum_{x^n \in A} \left[D(P_{Y^n|x^n} \| \gamma_A) + D(\gamma_A \| P_{Y^n|x^n}) \right] \quad (60)$$

$$= \frac{1}{|A|} \sum_{x^n \in A} \left[D(P_{Y^n|x^n} \| \mu_A) + D(\mu_A \| P_{Y^n|x^n}) \right] \quad (61)$$

$$= \frac{1}{|A|} \sum_{x^n \in A} \left[D(P_{Y^n|x^n} \| \mu_A) + D(\mu_A \| \gamma_A) + D(\gamma_A \| P_{Y^n|x^n}) \right]. \quad (62)$$

Note that $\frac{1}{|A|} \pi(A, A) + \frac{1}{|A^c|} \pi(A^c, A^c)$ is similar to a known clustering problem. In the clustering literature, the min-sum clustering problem [12] is minimizing the sum of all edges in each cluster. Using π , we can describe the binary min-sum clustering problem on Ω by minimizing $\pi(A, A) + \pi(A^c, A^c)$.

4.4. Another Application of the Geometric Mean

Using the geometric mean of measures, we can rewrite the clustering problem in a different form. Recall that $\mu_{A \oplus x^n} = \{\tilde{x}^n \oplus x^n : \tilde{x}^n \in A\}$. Then, we have:

$$\mathcal{D}(A) = \sum_{x^n \in A} D(P_{Y^n|x^n} \| \mu_A) \quad (63)$$

$$= \sum_{x^n \in A} D(P_{Y^n|0^n} \| \mu_{A \oplus x^n}). \quad (64)$$

Let $\tilde{\gamma}_A$ be the geometric mean of measures in the set $\{\mu_{A \oplus x^n} : x^n \in A\}$, i.e.,

$$\tilde{\gamma}_A(y^n) = \frac{1}{\tilde{s}_A} \left(\prod_{x^n \in A} \mu_{A \oplus x^n}(y^n) \right)^{1/|A|} \quad (65)$$

where:

$$\tilde{s}_A = \sum_{y^n} \left(\prod_{x^n \in A} \mu_{A \oplus x^n}(y^n) \right)^{1/|A|}. \quad (66)$$

Then, we have:

$$\mathcal{D}(A) = |A| D(P_{Y^n|0^n} \| \tilde{\gamma}_A) + |A| \log \frac{1}{\tilde{s}_A}. \quad (67)$$

Needless to say:

$$\mathcal{D}(A^c) = |A^c| D(P_{Y^n|0^n} \| \tilde{\gamma}_{A^c}) + |A^c| \log \frac{1}{\tilde{s}_{A^c}}. \quad (68)$$

The sum of the results is:

$$\mathcal{D}(A) + \mathcal{D}(A^c) = |A| D(P_{Y^n|0^n} \| \tilde{\gamma}_A) + |A^c| D(P_{Y^n|0^n} \| \tilde{\gamma}_{A^c}) + |A| \log \frac{1}{\tilde{s}_A} + |A^c| \log \frac{1}{\tilde{s}_{A^c}}. \quad (69)$$

This can be considered as a dual of Theorem 3.

Remark 7. Let $\Omega_{i0} = \{x^n : x_i = 0\}$, which is the candidate of the optimizer. Then,

$$\tilde{\gamma}_{\Omega_{i0}} = \tilde{\gamma}_{(\Omega_{i0})^c} = \mu_{\Omega_{i0}} \quad (70)$$

$$\tilde{s}_{\Omega_{i0}} = \tilde{s}_{(\Omega_{i0})^c} = 1. \quad (71)$$

5. Concluding Remarks

In this paper, we have proposed a number of different formulations of the most informative Boolean function conjecture. Most of them are based on the information geometric approach. Furthermore, we focused on the (normalized) geometric mean of measures that can simplify the problem formulation. More precisely, we showed that Conjecture 1 is true if and only if the maximum achieving f satisfies the following property: “the arithmetic and geometric mean of measures are the same for both $\{P_{Y^n|x^n} : x^n \in f^{-1}(0)\}$, as well as $\{P_{Y^n|x^n} : x^n \in f^{-1}(1)\}$.”

Funding: This work was supported by the Hongik University new faculty research support fund.

Conflicts of Interest: The author declares no conflict of interest.

Appendix A. Proof of Theorem 2

By the definition of mutual information, we have:

$$I(X; Y) - I(f(X); Y) = \mathbb{E} \left[\log \frac{P_{X,Y}(X, Y)}{P_X(X)P_Y(Y)} \right] - \mathbb{E} \left[\log \frac{P_{U,Y}(f(X), Y)}{P_U(f(X))P_Y(Y)} \right] \quad (A1)$$

$$= \mathbb{E} \left[\log \frac{P_{Y|X}(Y|X)}{P_{Y|U}(Y|f(X))} \right] \quad (A2)$$

$$= \mathbb{E} \left[\mathbb{E} \left[\log \frac{P_{Y|X}(Y|X)}{P_{Y|U}(Y|f(X))} \middle| X \right] \right] \quad (A3)$$

$$= \sum_x P_X(x) \mathbb{E} \left[\log \frac{P_{Y|X}(Y|X)}{P_{Y|U}(Y|f(X))} \middle| X = x \right] \quad (A4)$$

$$= \sum_x P_X(x) D(P_{Y|x} \| P_{Y|U}(\cdot | f(x))). \quad (A5)$$

This concludes the proof.

Appendix B. Proof of Theorem 4

Without loss of generality, we can assume that $x^n = 0^n$ and $\tilde{x}^n = 1^k 0^{n-k}$ where $k = d_H(x^n, \tilde{x}^n)$. Then, we have:

$$\begin{aligned} & D(P_{Y^n|x^n} \| P_{Y^n|\tilde{x}^n}) \\ &= D(P_{Y^k|x^k} \times P_{Y_{k+1}^n|x_{k+1}^n} \| P_{Y^k|\tilde{x}^k} \times P_{Y_{k+1}^n|\tilde{x}_{k+1}^n}) \end{aligned} \quad (A6)$$

$$= D(P_{Y^k|0^k} \times P_{Y_{k+1}^n|0_{k+1}^n} \| P_{Y^k|1^k} \times P_{Y_{k+1}^n|0_{k+1}^n}) \quad (A7)$$

$$= D(P_{Y^k|0^k} \| P_{Y^k|1^k}) + D(P_{Y_{k+1}^n|0_{k+1}^n} \| P_{Y_{k+1}^n|0_{k+1}^n}) \quad (A8)$$

$$= D(P_{Y^k|0^k} \| P_{Y^k|1^k}) \quad (A9)$$

Thus,

$$\begin{aligned} & D(P_{Y^k|0^k} \| P_{Y^k|1^k}) \\ &= kD(P_{Y|0} \| P_{Y|1}) \end{aligned} \quad (\text{A10})$$

$$= k \left(p \log \frac{p}{1-p} + (1-p) \log \frac{1-p}{p} \right) \quad (\text{A11})$$

$$= k(1-2p) \log \frac{1-p}{p}. \quad (\text{A12})$$

Since $d_H(x^n, \tilde{x}^n) = k$, this concludes the proof.

Appendix C. Proof of Theorem 5

The following lemma bounds the ratio between $Q_Y^n(y^n)$ and $Q_Y^n(\tilde{y}^n)$, which will be crucial in our argument.

Lemma A1. For $Q_{Y^n} \in \text{conv}\{P_{Y^n|X^n} | x^n \in \Omega\}$,

$$d_H(y^n, \tilde{y}^n) \cdot \log \left(\frac{p}{1-p} \right) \leq \log \frac{Q_{Y^n}(y^n)}{Q_{Y^n}(\tilde{y}^n)} \leq d_H(y^n, \tilde{y}^n) \cdot \log \left(\frac{1-p}{p} \right) \quad (\text{A13})$$

Proof. Without loss of generality, we can assume that $\tilde{y}^k = \tilde{y}^k$ and $y_{k+1}^n = \tilde{y}_{k+1}^n$.

$$\frac{Q_{Y^n}(y^n)}{Q_{Y^n}(\tilde{y}^k, y_{k+1}^n)} = \frac{\sum_{x^n} \pi(x^n) P_{Y^n|X^n}(y^n | x^n)}{\sum_{x^n} \pi(x^n) P_{Y^n|X^n}(\tilde{y}^k, y_{k+1}^n | x^n)} \quad (\text{A14})$$

$$\leq \left(\frac{1-p}{p} \right)^k \frac{\sum_{x^n} \pi(x^n) P_{Y^n|X^n}(y^n | x^n)}{\sum_{x^n} \pi(x^n) P_{Y^n|X^n}(y^n | x^n)} \quad (\text{A15})$$

$$= \left(\frac{1-p}{p} \right)^k. \quad (\text{A16})$$

Similarly, we can show that:

$$\frac{Q_{Y^n}(y^n)}{Q_{Y^n}(\tilde{y}^k, y_{k+1}^n)} \geq \left(\frac{p}{1-p} \right)^k. \quad (\text{A17})$$

This concludes the proof of the lemma. \square

Without loss of generality, we can assume that $x^n = 0^n$ and $\tilde{x}^n = 1^k 0^{n-k}$. Then, we have:

$$P_{Y^n|X^n}(y^n | \tilde{x}^n) = P_{Y^n|X^n}(y^n | 1^k 0^{n-k}) \quad (\text{A18})$$

$$= P_{Y^n|X^n}(\tilde{y}^k, y_{k+1}^n | 0^n) \quad (\text{A19})$$

where $\tilde{y}_i = 1 - y_i$. Thus, we have:

$$\begin{aligned} & D(P_{Y^n|X^n=x^n} \| Q_{Y^n}) - D(P_{Y^n|X^n=\tilde{x}^n} \| Q_{Y^n}) \\ &= \mathbb{E} \left[\log \frac{P_{Y^n|X^n}(Y^n | 0^n)}{Q_{Y^n}(Y^n)} \right] - \mathbb{E} \left[\log \frac{P_{Y^n|X^n}(\tilde{Y}^k, Y_{k+1}^n | 0^n)}{Q_{Y^n}(Y^n)} \right] \end{aligned} \quad (\text{A20})$$

$$= \mathbb{E} \left[\log \frac{P_{Y^n|X^n}(Y^n|0^n)}{Q_{Y^n}(Y^n)} \right] - \mathbb{E} \left[\log \frac{P_{Y^n|X^n}(Y^n|0^n)}{Q_{Y^n}(\bar{Y}^k, Y_{k+1}^n)} \right] \quad (\text{A21})$$

$$= \mathbb{E} \left[\log \frac{Q_{Y^n}(Y^n)}{Q_{Y^n}(\bar{Y}^k, Y_{k+1}^n)} \right]. \quad (\text{A22})$$

Note that two expectations in (A20) are under different distributions; on the other hand, expectations in (A21) and the following equations are under the same distribution $P_{Y^n|X^n=0^n}$.

The above expectation can be written as follows.

$$D(P_{Y^n|X^n=x^n} \| Q_{Y^n}) - D(P_{Y^n|X^n=\bar{x}^n} \| Q_{Y^n}) \\ = \sum_{y^n} P_{Y^n|X^n}(y^n|0^n) \log \frac{Q_{Y^n}(y^n)}{Q_{Y^n}(\bar{y}^k, y_{k+1}^n)} \quad (\text{A23})$$

$$= \sum_{y_2^n} P_{Y^n|X^n}(0, y_2^n|0^n) \log \frac{Q_{Y^n}(0, y_2^n)}{Q_{Y^n}(1, \bar{y}_2^k, y_{k+1}^n)} \\ + \sum_{y_2^n} P_{Y^n|X^n}(1, y_2^n|0^n) \log \frac{Q_{Y^n}(1, y_2^n)}{Q_{Y^n}(0, \bar{y}_2^k, y_{k+1}^n)} \quad (\text{A24})$$

$$= \sum_{y_2^n} P_{Y^n|X^n}(0, y_2^n|0^n) \log \frac{Q_{Y^n}(0, y_2^n)}{Q_{Y^n}(1, \bar{y}_2^k, y_{k+1}^n)} \\ + \sum_{y_2^n} P_{Y^n|X^n}(1, \bar{y}_2^k, y_{k+1}^n|0^n) \log \frac{Q_{Y^n}(1, \bar{y}_2^k, y_{k+1}^n)}{Q_{Y^n}(0, y_2^n)} \quad (\text{A25})$$

$$= \sum_{y_2^n} \left((1-p)P_{Y_2^n|X_2^n}(y_2^n|0^{n-1}) - pP_{Y_2^n|X_2^n}(\bar{y}_2^k, y_{k+1}^n|0^{n-1}) \right) \log \frac{Q_{Y^n}(0, y_2^n)}{Q_{Y^n}(1, \bar{y}_2^k, y_{k+1}^n)} \quad (\text{A26})$$

$$\leq \sum_{y_2^n} \left| (1-p)P_{Y_2^n|X_2^n}(y_2^n|0^{n-1}) - pP_{Y_2^n|X_2^n}(\bar{y}_2^k, y_{k+1}^n|0^{n-1}) \right| k \log \frac{1-p}{p} \quad (\text{A27})$$

$$= \sum_{y_2^k} \left| (1-p)P_{Y_2^k|X_2^k}(y_2^k|0^{k-1}) - pP_{Y_2^k|X_2^k}(\bar{y}_2^k|0^{k-1}) \right| k \log \frac{1-p}{p} \quad (\text{A28})$$

where (A27) is because of Lemma A1.

Finally,

$$\sum_{y_2^k} \left| (1-p)P_{Y_2^k|X_2^k}(y_2^k|0^{k-1}) - pP_{Y_2^k|X_2^k}(\bar{y}_2^k|0^{k-1}) \right| \\ \leq \sum_{i=0}^{k-1} \binom{k-1}{i} \left| (1-p)p^i(1-p)^{k-1-i} - p \cdot p^{k-1-i}(1-p)^i \right| \quad (\text{A29})$$

$$\leq (1-p)^k - p^k. \quad (\text{A30})$$

This concludes the proof.

Appendix D. Proof of Theorem 6

From the first assumption $\mu_A = \gamma_A$, we have:

$$\sum_{y^{n-1}} \mu_A(y^{n-1}0) = \sum_{y^{n-1}} \frac{1}{|A|} \sum_{x^n \in A} P_{Y^n|X^n}(y^{n-1}0|x^n) \quad (\text{A31})$$

$$= \sum_{y^{n-1}} \frac{1}{|A|} \sum_{x^n \in A_{n0}} P_{Y^n|X^n}(y^{n-1}0|x^n) + \sum_{y^{n-1}} \frac{1}{|A|} \sum_{x^n \in A_{n1}} P_{Y^n|X^n}(y^{n-1}0|x^n) \quad (\text{A32})$$

$$= \frac{1-p}{|A|} \sum_{x^n \in A_{n0}} \sum_{y^{n-1}} P_{Y^{n-1}|X^{n-1}}(y^{n-1}|x^{n-1}) + \frac{p}{|A|} \sum_{x^n \in A_{n1}} \sum_{y^{n-1}} P_{Y^{n-1}|X^{n-1}}(y^{n-1}|x^{n-1}) \quad (\text{A33})$$

$$= (1-p) \frac{|A_{n0}|}{|A|} + p \frac{|A_{n1}|}{|A|}. \quad (\text{A34})$$

Clearly, we can get the following result in a similar manner.

$$\sum_{y^{n-1}} \mu_A(y^{n-1}1) = p \frac{|A_{n0}|}{|A|} + (1-p) \frac{|A_{n1}|}{|A|}. \quad (\text{A35})$$

The ratio of those two is given by:

$$\frac{\sum_{y^{n-1}} \mu_A(y^{n-1}1)}{\sum_{y^{n-1}} \mu_A(y^{n-1}0)} = \frac{p \frac{|A_{n0}|}{|A|} + (1-p) \frac{|A_{n1}|}{|A|}}{(1-p) \frac{|A_{n0}|}{|A|} + p \frac{|A_{n1}|}{|A|}}. \quad (\text{A36})$$

On the other hand, we can also marginalize γ_A :

$$\sum_{y^{n-1}} \gamma_A(y^{n-1}0) = \frac{1}{s_A} \sum_{y^{n-1}} \left(\prod_{x^n \in A} P_{Y^n|X^n}(y^{n-1}0|x^n) \right)^{1/|A|} \quad (\text{A37})$$

$$= \frac{1}{s_A} \sum_{y^{n-1}} \left(\prod_{x^n \in A_{n0}} (1-p) P_{Y^{n-1}|X^{n-1}}(y^{n-1}|x^{n-1}) \prod_{x^n \in A_{n1}} p P_{Y^{n-1}|X^{n-1}}(y^{n-1}|x^{n-1}) \right)^{1/|A|} \quad (\text{A38})$$

$$= \frac{1}{s_A} \sum_{y^{n-1}} \left((1-p)^{|A_{n0}|} p^{|A_{n1}|} \prod_{x^n \in A} P_{Y^{n-1}|X^{n-1}}(y^{n-1}|x^{n-1}) \right)^{1/|A|} \quad (\text{A39})$$

$$= \frac{1}{s_A} (1-p)^{|A_{n0}|/|A|} p^{|A_{n1}|/|A|} \sum_{y^{n-1}} \left(\prod_{x^n \in A} P_{Y^{n-1}|X^{n-1}}(y^{n-1}|x^{n-1}) \right)^{1/|A|}. \quad (\text{A40})$$

Similarly, we have:

$$\sum_{y^{n-1}} \gamma_A(y^{n-1}1) = \frac{1}{s_A} p^{|A_{n0}|/|A|} (1-p)^{|A_{n1}|/|A|} \sum_{y^{n-1}} \left(\prod_{x^n \in A} P_{Y^{n-1}|X^{n-1}}(y^{n-1}|x^{n-1}) \right)^{1/|A|}. \quad (\text{A41})$$

Thus, the ratio is given by:

$$\frac{\sum_{y^{n-1}} \gamma_A(y^{n-1}1)}{\sum_{y^{n-1}} \gamma_A(y^{n-1}0)} = \frac{p^{|A_{n0}|/|A|} (1-p)^{|A_{n1}|/|A|}}{(1-p)^{|A_{n0}|/|A|} p^{|A_{n1}|/|A|}} \quad (\text{A42})$$

Since $\mu_A = \gamma_A$, both ratios should be the same. Let $x = |A_{n0}|/|A|$, which implies $|A_{n1}|/|A| = 1 - x$. Then, we have:

$$\frac{px + (1-p)(1-x)}{(1-p)x + p(1-x)} = \frac{p^x(1-p)^{1-x}}{(1-p)^x p^{1-x}}. \quad (\text{A43})$$

If we let $y = \frac{p}{1-p}$, then the above equation can be further simplified to:

$$\frac{xy + (1-x)}{x + y(1-x)} = y^{2x-1}. \quad (\text{A44})$$

Lemma A2. For fixed $0 < y < 1$, the only solutions of the above equation are $x = 0, \frac{1}{2}, 1$.

Proof. It is clear that $x = 0, \frac{1}{2}, 1$ comprise the solution of the following equation.

$$\frac{xy + (1-x)}{x + y(1-x)} = y^{2x-1}. \quad (\text{A45})$$

It is enough to show that:

$$g_y(x) = \log(xy + (1-x)) - \log(x + y(1-x)) - (2x-1)\log y = 0 \quad (\text{A46})$$

can have up to three solutions. Consider the derivative $\frac{\partial}{\partial x} g_y(x) = 0$,

$$\frac{\partial}{\partial x} g_y(x) = \frac{y-1}{xy+1-x} - \frac{1-y}{x+y-xy} - 2\log y = 0 \quad (\text{A47})$$

which is equivalent to:

$$2(xy + 1 - x)(x + y - xy) \log y = (1 + y)(y - 1). \quad (\text{A48})$$

It is a quadratic equation, and therefore, $\frac{\partial}{\partial x} g_y(x) = 0$ can have up to two solutions. Thus, $g_y(x) = 0$ can have up to three solutions. \square

This implies that $|A_{n0}| = 0, |A|/2$ or $|A|$. It is clear that the above result holds for all i and A^c , i.e., $|A_{i0}|$ is either 0, $|A|/2$ or $|A|$, and $|A_{i0}^c|$ is either 0, $|A^c|/2$ or $|A^c|$. These cardinalities should satisfy the following equations:

$$|A_{i0}| + |A_{i1}| = |A| \quad (\text{A49})$$

$$|A_{i0}| + |A_{i0}^c| = 2^{n-1} \quad (\text{A50})$$

$$|A_{i0}^c| + |A_{i1}^c| = |A^c| \quad (\text{A51})$$

$$|A_{i1}| + |A_{i1}^c| = 2^{n-1} \quad (\text{A52})$$

for all $1 \leq i \leq n$. Since \mathcal{I} -compressedness implies $A_{i1} \subseteq A_{i0}$, we have $|A_{i1}| \leq |A_{i0}|$. Thus, $|A_{i0}|$ should be either $|A|/2$ or $|A|$ for all i . If $|A_{i0}| = |A|$ for some i , then $|A_{i1}| = 0$. Since $|A_{i1}^c|$ is either 0, $|A^c|/2$ or $|A^c|$, but $A^c \neq \emptyset, \Omega$, we have $|A_{i1}^c| = 2^{n-1}$. Thus, $A = A_{i0}$ and $|A| = 2^{n-1}$, which implies $A = \Omega_{i0}$.

On the other hand, assume that $|A_{i0}| = |A_{i1}| = |A|/2$ for all i . Since A is \mathcal{I} -compressed, $x^{i-1}1x_{i+1}^n \in A_{i1}$ implies $x^{i-1}0x_{i+1}^n \in A_{i0}$. However, we have $|A_{i0}| = |A_{i1}|$, and therefore:

$$x^{i-1}1x_{i+1}^n \in A_{i1} \Leftrightarrow x^{i-1}0x_{i+1}^n \in A_{i0} \quad (\text{A53})$$

and equivalently,

$$x^{i-1}1x_{i+1}^n \in A \Leftrightarrow x^{i-1}0x_{i+1}^n \in A \quad (\text{A54})$$

for all i . It can only be true when $A = \emptyset$ or Ω , which contradicts our original assumption. This concludes the proof.

References

1. Courtade, T.A.; Kumar, G.R. Which Boolean functions maximize mutual information on noisy inputs? *IEEE Trans. Inf. Theory* **2014**, *60*, 4515–4525. [[CrossRef](#)]
2. Pichler, G.; Matz, G.; Piantanida, P. A tight upper bound on the mutual information of two Boolean functions. In Proceedings of the 2016 IEEE Information Theory Workshop (ITW), Cambridge, UK, 11–14 September 2016; pp. 16–20.
3. Ordentlich, O.; Shayevitz, O.; Weinstein, O. An improved upper bound for the most informative Boolean function conjecture. In Proceedings of the 2016 IEEE International Symposium on Information Theory (ISIT), Barcelona, Spain, 10–15 July 2016; pp. 500–504.
4. Weinberger, N.; Shayevitz, O. On the optimal Boolean function for prediction under quadratic loss. *IEEE Trans. Inf. Theory* **2017**, *63*, 4202–4217. [[CrossRef](#)]
5. Huleihel, W.; Ordentlich, O. How to quantize n outputs of a binary symmetric channel to $n-1$ bits? In Proceedings of the 2017 IEEE International Symposium on Information Theory (ISIT), Aachen, Germany, 25–30 June 2017; pp. 91–95.
6. Nazer, B.; Ordentlich, O.; Polyanskiy, Y. Information-distilling quantizers. In Proceedings of the 2017 IEEE International Symposium on Information Theory (ISIT), Aachen, Germany, 25–30 June 2017; pp. 96–100.
7. Lin, J. Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory* **1991**, *37*, 145–151. [[CrossRef](#)]
8. Topsøe, F. An information theoretical identity and a problem involving capacity. *Stud. Sci. Math. Hung.* **1967**, *2*, 291–292.
9. Banerjee, A.; Merugu, S.; Dhillon, I.S.; Ghosh, J. Clustering with Bregman divergences. *J Mach. Learn. Res.* **2005**, *6*, 1705–1749.
10. Nielsen, F. An information-geometric characterization of Chernoff information. *IEEE Signal Process. Lett.* **2013**, *20*, 269–272. [[CrossRef](#)]
11. Nielsen, F.; Boltz, S. The burbea-rao and bhattacharyya centroids. *IEEE Trans. Inf. Theory* **2011**, *57*, 5455–5466. [[CrossRef](#)]
12. Guttmann-Beck, N.; Hassin, R. Approximation algorithms for min-sum p -clustering. *Discrete Appl. Math.* **1998**, *89*, 125–142. [[CrossRef](#)]



© 2018 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).