Which Boolean Functions Maximize Mutual Information on Noisy Inputs?

Thomas A. Courtade, *Member, IEEE,* and Gowtham R. Kumar, *Student Member, IEEE*

***Abstract*— We pose a simply stated conjecture regarding the maximum mutual information a Boolean function can reveal about noisy inputs. Specifically, let X^n be independent identically distributed Bernoulli(1/2), and let Y^n be the result of passing X^n through a memoryless binary symmetric channel with crossover probability α. For any Boolean function $b : \{0,1\}^n \to \{0,1\}$, we conjecture that $I(b(X^n); Y^n) \leq 1 - H(\alpha)$. While the conjecture remains open, we provide substantial evidence supporting its validity. Connections are also made to discrete isoperimetric inequalities.**

***Index Terms*— Boolean functions, mutual information, extremal inequality, isoperimetric inequality.**

I. Introduction

THIS paper is inspired by the following conjecture:

Conjecture 1: Let $X^n = (X_1, \dots, X_n)$ be a sequence of n i.i.d. Bernoulli(1/2) random variables, and let Y^n be the result of passing X^n through a memoryless binary symmetric channel with crossover probability α. For any Boolean function $b : \{0,1\}^n \to \{0,1\}$, we have

$$I(b(X^n); Y^n) \leq 1 - H(\alpha). \tag{1}$$

At first sight, Conjecture 1 might appear to be no more than a simple exercise. However, over the course of this paper, we hope to convince the reader that the conjecture is much deeper than it appears. Despite its apparent simplicity, standard information-theoretical manipulations appear incapable of establishing (1).

To the present authors, Conjecture 1 represents the simplest, nontrivial embodiment of Boolean functions in an information-theoretic context. In words, Conjecture 1 asks: "*What is the most significant bit that X^n can provide about Y^n?*"

Despite their fundamental roles in computer science and digital computation, Boolean functions have received relatively little attention from the information theory community. The recent work [2] is perhaps most relevant to our Conjecture 1 and provides compelling motivation for its study. In [2], Klotz et al. prove that for n and $\mathscr{P}\{b(X^n) = 0\} \geq 1/2$ fixed, $I(b(X^n); X_1)$ is maximized by functions b which satisfy $b(X^n) = 0$ whenever $X_1 = 0$ (i.e., when b is *canalizing* in X_1). Their motivation for considering this problem comes from computational biology, where Boolean functions are used to model dependencies in various regulatory networks. We encourage the reader to refer to [2], [3] and the references therein for further information.

By employing Fourier-analytic techniques, we can obtain the following result:

Theorem 1: If $b(X^n)$ is equiprobable, then

$$\sum_{i=1}^{n} I(b(X^n); Y_i) \leq 1 - H(\alpha). \tag{2}$$

While Theorem 1 is weaker[1] than Conjecture 1, it is striking nonetheless. Indeed, it states that the sum of the n mutual information terms in (2) cannot exceed $1 - H(\alpha)$, regardless of how large n is. Unfortunately, our Fourier-analytic approach appears incapable of establishing Conjecture 1.

In addition to the clear connection to [2], Conjecture 1 is also related to the *Information Bottleneck Method* [4], which attempts to solve the optimization problem

$$\min_{p(u|x^n)} I(X^n; U) - \lambda I(Y^n; U). \tag{3}$$

For a given $\lambda > 0$, the optimizing U intuitively provides the best tradeoff between the accuracy of describing Y^n and the descriptive complexity of U. In our setting, $b(X^n)$ plays the role of U, and we constrain the descriptive complexity to be at most one bit. While U is allowed to be a stochastic function of X^n in (3), it is straightforward to show that randomized Boolean functions do not yield a higher mutual information in our setting (see Appendix A). Thus, expressing Conjecture 1 in terms of deterministic Boolean functions comes without loss of generality.

A more concrete example comes in the context of gambling. To this end, suppose Y^n is a simple model for a market of n stocks, where each stock increases or decreases in value with probability 1/2, independent of all other stocks, and we can buy *put* and *call* double-or-nothing options on each outcome. If an oracle has access to side information X^n, and we are allowed to ask *one* yes/no question of the oracle,

Manuscript received November 7, 2013; revised April 14, 2014; accepted May 7, 2014. Date of publication May 29, 2014, date of current version July 10, 2014. This work was supported in part by the National Science Foundation Center for Science of Information under Grant CCF-0939370 and in part by the Air Force under Grant FA9550-10-1-0124. This paper was presented at the 2013 IEEE International Symposium on Information Theory [1].

T. A. Courtade is with the University of California at Berkeley, Berkeley, CA 94720 USA (e-mail: courtade@eecs.berkeley.edu).

G. R. Kumar is with Stanford University, Stanford, CA 94305 USA (e-mail: gowthamr@stanford.edu).

Communicated by J. Körner, Associate Editor for Shannon Theory.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2014.2326877

[1] To see that (2) is weaker than Conjecture 1, note that (1) is equivalent to $\sum_{i=1}^{n} I\left(Y^{i-1}, b(X^n); Y_i\right) \leq 1 - H(\alpha)$ by independence of the Y_i's.

which question should we ask to maximize the rate at which our wealth grows? The validity of Conjecture 1 would imply that we should only concern ourselves with the performance of a single stock, say Y_1. This is readily seen as a consequence of known results on gambling with side information [5, Th. 16.4.2], since putting $b(X^n) = X_1$ yields

$$I(b(X^n); Y^n) = I(X_1; Y^n) = I(X_1; Y_1) = 1 - H(\alpha), \quad (4)$$

hence the conjectured upper bound (1) is achieved and represents the maximum attainable increase in doubling rate. On that note, Erkip applied Mrs. Gerber's lemma [6] to show the bound

$$I(b(X^n); Y^n) \leq (1 - 2\alpha)^2, \quad (5)$$

in her Ph.D. thesis [7, Th. 5]. However, (5) is strictly weaker than Conjecture 1.

Finally, we point out that (1) is related in spirit to the notion of *average sensitivity* of Boolean functions. This topic has received a great deal of attention in the computer science literature (see the survey [8], and references therein). To see the connection to sensitivity, note that (1) can be rewritten as

$$H(b(X^n)|Y^n) \geq H(b(X^n)) - 1 + H(\alpha). \quad (6)$$

For fixed $\mathscr{P}\{b(X^n) = 0\}$, the right hand side of (6) is constant. Hence, our conjecture essentially lower bounds the output uncertainty of Boolean functions with respect to noisy inputs. A longstanding conjecture that is similar in spirit to our Conjecture 1 is the *Entropy-Influence Conjecture*. Though not necessarily information-theoretic in nature, the entropy-influence conjecture postulates that the average sensitivity of a Boolean function can be lower bounded (up to a constant factor) by the entropy of its squared Fourier coefficients. Although the conjectures appear to be only superficially similar, we refer the interested reader to [9] for details and a precise statement.

The remainder of this paper is organized as follows. Section II provides a summary of the main results and their implications. Primarily, it focuses on a refinement of Conjecture 1 into two "sub-conjectures" and provides evidence supporting their validity. Section III contains the proofs of the main results, and Section IV delivers our concluding remarks.

II. Main Results and Implications

A. Notation and Definitions

Before proceeding, we establish the basic notation and assumptions which will be used throughout the paper. Let $X^n = (X_1, X_2, \dots, X_n)$ be a sequence of n i.i.d. Bernoulli$(1/2)$ random variables, and let $Z^n = (Z_1, Z_2, \dots, Z_n)$ be a sequence of n i.i.d. Bernoulli(α) random variables independent of X^n, with $0 \leq \alpha \leq 1/2$. Define $Y^n = X^n \oplus Z^n$, where "$\oplus$" denotes coordinate-wise XOR. In other words, Y^n is the result of passing X^n through a memoryless binary symmetric channel with crossover probability α.

Throughout, we let $\Omega = \{0, 1\}$, $\Omega_n = \{0, 1\}^n$, and consider Boolean functions $b : \Omega_n \to \Omega$. For a given function b, the notation $b^{-1}(0)$ will be used to denote the set $\{x^n \in \Omega_n : b(x^n) = 0\}$.

For a scalar $p \in [0, 1]$, $H(p)$ will be used to denote the binary entropy function $p \log \frac{1}{p} + (1 - p) \log \frac{1}{1-p}$. On the other hand, if W is a random variable, we write $H(W)$ for the Shannon entropy of W. The different usages of $H(\cdot)$ will be clear from context. All logarithms are assumed to be base-2 unless stated otherwise.

Definition 1: The lexicographical ordering $\prec_L$ on $\{0, 1\}^k$ is defined as follows: $x^k \prec_L \tilde{x}^k$ iff $x_j < \tilde{x}_j$ for some j and $x_i = \tilde{x}_i$ for all $i < j$.

For example, if $k = 3$, we have $000 \prec_L 001 \prec_L 010 \prec_L 011 \prec_L 100 \prec_L 101 \prec_L 110 \prec_L 111$.

Definition 2: We define $L_k(M)$ to be the initial segment of size M in the lexicographical ordering on $\{0, 1\}^k$. For example, $L_3(4) = \{000, 001, 010, 011\}$.

For a function $b : \Omega_n \to \Omega$, we say that "b is *lex*" when $b^{-1}(0) = L_n(|b^{-1}(0)|)$. In other words, b is lex when it maps an initial segment of the lexicographical order to 0, and the complement segment to 1.

B. Refinement of Conjecture 1

Instead of dealing with Conjecture 1 directly, consider the following two conjectures:

Conjecture 2: For a given n and fixed bias $\mathscr{P}\{b(X^n) = 0\}$ satisfying $H(\mathscr{P}\{b(X^n) = 0\}) \geq 1 - H(\alpha)$, the conditional entropy $H(b(X^n)|Y^n)$ is minimized when b is lex.

Conjecture 3: If $b : \Omega_n \to \Omega$ is lex, then

$$H(b(X^n)|Y^n) \geq H(b(X^n))H(\alpha). \quad (7)$$

Clearly, Conjecture 1 would follow as a corollary if Conjectures 2 and 3 were valid. Though it might seem counterproductive to try to prove a stronger result by splitting Conjecture 1 into two different conjectures, this appears to be the only way to gain traction on Conjecture 1. Intuitively, Conjecture 2 concerns the *structure* of maximally-informative Boolean functions, while Conjecture 3 handles the *inequality component* of Conjecture 1. This intuition should become more clear to the reader over the course of this paper.

We briefly comment that our reference to Conjecture 3 as a "conjecture" is perhaps too modest. Indeed, for any given α, we have a simple recursive algorithm capable of proving (7) for all n. With computer assistance, our algorithm has verified Conjecture 3 for α ranging from 0 to $1/2$ in increments of 0.001 (refer to Theorem 4 and the following discussion in Section II-D for details). Nevertheless, given the computer-assisted nature of our method for verifying (7), we find it appropriate to call it a conjecture.

In the following two subsections, we elaborate on Conjectures 2 and 3 and our corresponding results.

C. Conjecture 2 and Edge-Isoperimetry

Conjecture 2 is reminiscent of a classical theorem in discrete mathematics originally due to Harper [10] that gives an exact edge-isoperimetric inequality for the hypercube. To state the theorem, we need a few basic notations. Let Q_n be the n-dimensional hypercube, and let $V(Q_n) = \Omega_n$ be its set of vertices. For $S \subseteq V(Q_n)$, the edge boundary $\partial(S)$ is the set

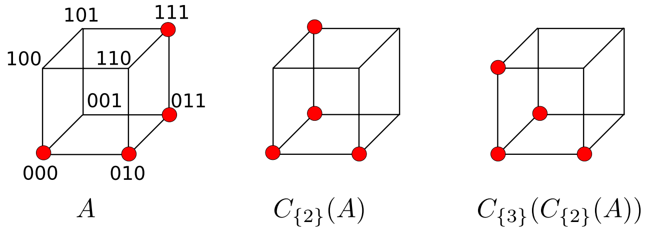

Fig. 1. Successive one-dimensional compressions applied to the set $A = \{000, 010, 011, 111\}$, elements of which are represented by balls on the three-dimensional hypercube. (Vertices are only labeled on the leftmost cube for cleaner presentation.)

TABLE I

Reduction in Number of Candidate Boolean Functions to be Considered for Verification of Conjecture 2

n	$\lvert\mathcal{S}_n\rvert$	$\lvert\mathcal{B}_n\rvert$
2	5	16
3	10	256
4	27	65,536
5	119	4.3×10^9
6	1173	1.8×10^{19}
7	44,315	3.4×10^{38}

of edges one has to delete to disconnect S from any vertex not in S.

Theorem 2 (Harper's Edge-Isoperimetric Inequality): For $S \subseteq V(Q_n)$ with $|S| = k$, we have $|\partial(S)| \geq |\partial(L_n(k))|$.

We comment on the relationship between edge-isoperimetry and our conjecture in detail in Appendix B.

The simplest proofs of Theorem 2 rely on so-called *compression operators*, popularized by Bollobás and Leader [11]. These compression operators turn out to be useful in making progress towards Conjecture 2, so we introduce them now.

Let $\mathcal{I}$ be subset of $\{1, 2, \ldots, n\}$ of cardinality k. To be concrete, let $\mathcal{I} = \{i_1, i_2, \ldots, i_k\}$, where $i_1 < i_2 < \cdots < i_k$. Following the notation of [11], for a set $B \subseteq \Omega_n$ and x^n having $x_i = 0$ for all $i \in \mathcal{I}$, we define the $\mathcal{I}$-section of B at x^n by

$$B_{\mathcal{I}}(x^n) = \left\{ z^k : y^n \in B, y_i = \left\{ \begin{array}{ll} z_j & \text{if } i = i_j \in \mathcal{I} \\ x_i & \text{otherwise} \end{array} \right. \right\}. \quad (8)$$

A few examples will help clarify (8). For instance, if $B = \{000, 001, 011, 101\}$, a few examples of $\mathcal{I}$-sections at different $x^3 \in \Omega_3$ are given by:

$$B_{\{1\}}(001) = \{0, 1\}, \quad (9)$$
$$B_{\{2\}}(100) = \emptyset, \quad (10)$$
$$B_{\{1,2\}}(000) = \{00\}, \quad (11)$$
$$B_{\{1,2\}}(001) = \{00, 01, 10\}. \quad (12)$$

The $\mathcal{I}$-compression of B, $C_{\mathcal{I}}(B)$, is defined in terms of its $\mathcal{I}$-sections:

$$(C_{\mathcal{I}}(B))_{\mathcal{I}}(x^n) = L_k\left(|B_{\mathcal{I}}(x^n)|\right) \quad \text{for all } x^n.$$

In other words, $C_{\mathcal{I}}$ replaces each $\mathcal{I}$-section of B with an initial segment of the lexicographical order. We say that B is $\mathcal{I}$-compressed if $C_{\mathcal{I}}(B) = B$. Note that $C_{\mathcal{I}}(B)$ is always $\mathcal{I}$-compressed.

Continuing the above example of $B = \{000, 001, 011, 101\}$, example $\mathcal{I}$-compressions are given by:

$$C_{\{1\}}(B) = \{000, 001, 011, 101\}, \quad (13)$$
$$C_{\{2\}}(B) = \{000, 001, 011, 101\}, \quad (14)$$
$$C_{\{2,3\}}(B) = \{000, 001, 010, 100\}, \quad (15)$$
$$C_{\{1,3\}}(B) = \{000, 001, 010, 100\}. \quad (16)$$

To further illustrate the action of compression operators, successive compressions on the set $A = \{000, 010, 011, 111\}$ are illustrated in Figure 1.

We pause to make two important observations. First, $\mathcal{I}$-compression preserves the size of the set on which it operates. That is, $|C_{\mathcal{I}}(B)| = |B|$. Second, if B is $\mathcal{I}$-compressed, then it is also $\mathcal{J}$-compressed for all $\mathcal{J} \subset \mathcal{I}$.

The following theorem states that when $|\mathcal{I}| = 2$, applying an $\mathcal{I}$-compression to $b^{-1}(0)$ does not decrease the mutual information $I(b(X^n); Y^n)$. Thus, compression provides a method of modifying functions in a manner that does not adversely affect the mutual information $I(b(X^n); Y^n)$.

Theorem 3: Let $b : \Omega_n \to \Omega$ and let $\mathcal{I} \subseteq \{1, \ldots, n\}$ satisfy $|\mathcal{I}| = 2$. If $\hat{b} : \Omega_n \to \Omega$ is defined by its preimage $\hat{b}^{-1}(0) = C_{\mathcal{I}}(b^{-1}(0))$, then $I(\hat{b}(X^n); Y^n) \geq I(b(X^n); Y^n)$.

By definition, if $C_{\mathcal{I}}(\cdot)$ changes an element of $b^{-1}(0)$, it moves it *lower* in the lexicographical ordering on Ω_n. Therefore, a repeated application of Theorem 3 for different subsets $\mathcal{I}$ of cardinality 2 is guaranteed to terminate at a function $\hat{b}$ which is $\mathcal{I}$-compressed for all $\mathcal{I}$ with $|\mathcal{I}| \leq 2$. Hence, we have the following corollary.

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

The implications of Theorem 3 and its corollary are twofold. First, it allows the verification of Conjecture 2 for modest values of n. Indeed, we have numerically validated Conjectures 1 and 2 for $n \leq 7$ by enumerating the functions in $\mathcal{S}_n$ (see Appendix C for details) and evaluating $I(b(X^n); Y^n)$ for $b \in \mathcal{S}_n$. To appreciate the reduction afforded by Corollary 1, define $\mathcal{B}_n$ to be the set of all 2^{2^n} Boolean functions on n inputs. A comparison between $|\mathcal{S}_n|$ and $|\mathcal{B}_n|$ is given in Table I. As illustrated by Table I, a brute-force attempt to validate Conjecture 1 for $n \leq 7$ would be computationally intractable, yet Theorem 3 allows us to do so by considering only those functions in $\mathcal{S}_n$. In fact, one could potentially leverage Theorem 3 to exhaustively verify Conjecture 1 for modestly larger values of n (e.g., $n = 8, 9$), though we have not done so.

Second, Theorem 3 reinforces the intuition behind Conjecture 2. As we noted above, if $C_{\mathcal{I}}(\cdot)$ changes an element of $b^{-1}(0)$, it moves it *lower* in the lexicographical ordering on Ω_n. Thus, roughly speaking, applying $\mathcal{I}$-compression to $b^{-1}(0)$ yields a function $\hat{b}$ which is (i) closer to being lex, and (ii) for $|\mathcal{I}| \leq 2$ satisfies $H(\hat{b}(X^n)|Y^n) \leq H(b(X^n)|Y^n)$.

Ideally, Theorem 3 should generalize to include $\mathcal{I}$-compressions for $|\mathcal{I}| > 2$. Indeed, if we could take

$|\mathcal{I}| = n$, Conjecture 2 would be proved. However, we have found counterexamples where compression *increases* $H(b(X^n)|Y^n)$ for $|\mathcal{I}| > 2$ (but still reduces $H(b(X^n)|Y^n)$ for $|\mathcal{I}| = n$). One such counterexample is given as follows. Let $n = 6$ and suppose b_1 and b_2 are functions defined by

$$b_1^{-1}(0) = \begin{cases} 000000 \\ 000001 \\ 000010 \\ 000100 \\ 001000 \\ 010000 \\ 100000 \end{cases} \quad b_2^{-1}(0) = \begin{cases} 000000 \\ 000001 \\ 000010 \\ 000011 \\ 001000 \\ 010000 \\ 100000 \end{cases} \tag{17}$$

where the respective rows correspond to vectors composing $b_1^{-1}(0)$ and $b_2^{-1}(0)$. Note that b_2 is obtained from b_1 by a compression along the last three dimensions. Specifically, $b_2^{-1}(0) = C_{\{4,5,6\}}(b_1^{-1}(0))$. However, for $\alpha = 0.1$, computation reveals that $I(b_1(X^n); Y^n) = 0.2186$, while $I(b_2(X^n); Y^n) = 0.2173$. Therefore, we see that $\mathcal{I}$-compressions do not always improve mutual information for $|\mathcal{I}| \geq 3$. Nevertheless, letting b_L be the lex function for $n = 6$ and $|b_L^{-1}(0)| = 7$, we can compute $I(b_L(X^n); Y^n) = 0.2283$ for $\alpha = 0.1$, supporting Conjecture 2. In fact, as discussed above, Theorem 3 permits us to verify that 0.2283 bits is the maximum mutual information attainable among all functions b with $|b^{-1}(0)| = 7$ for $n = 6$, $\alpha = 0.1$.

Remark 1: From the above counterexample, it appears that the threshold function[2] *behaves like a local maximum in terms of mutual information provided about Y^n.*

Remark 2: A function $b : \Omega_n \to \Omega$ is said to be monotone *if $b(x^n) \leq b(\tilde{x}^n)$ whenever $x_i \leq \tilde{x}_i$ for $1 \leq i \leq n$. Since monotone Boolean functions are precisely those functions which are $\mathcal{I}$-compressed for all $\mathcal{I}$ with $|\mathcal{I}| = 1$, Corollary 1 implies that $I(b(X^n); Y^n)$ is maximized by a monotone function. Stronger still, Theorem 3 asserts that we need only consider an exponentially small class of monotone functions known as* regular *functions (see [23, Chapter 8] for a precise definition).*

D. A Computer-Assisted Proof of Conjecture 3 for any given α

Now, we turn toward establishing Conjecture 3. Unless otherwise specified, all Boolean functions in this subsection are assumed to be lex.

Define $f(x) = -x \log x$. Note that if b is lex, then so is the negation of b up to a relabeling of inputs. Specifically, $\tilde{b}(x^n) \triangleq 1 - b(\neg x^n)$ is lex if b is lex, where $\neg x^n$ is the negation of x^n. Therefore, to prove (7), it is sufficient to prove

$$\mathbb{E}_{Y^n} f\left(\mathscr{P}\{b(X^n) = 0 | Y^n\}\right) \geq f\left(\mathscr{P}\{b(X^n) = 0\}\right) H(\alpha). \tag{18}$$

To simplify notation, for a dyadic rational $p = k/2^n$, define

$$T_\alpha(p) \triangleq \mathbb{E}_{Y^n} f\left(\mathscr{P}\{b(X^n) = 0 | Y^n\}\right), \tag{19}$$

where b is the unique lex function on n inputs with $\mathscr{P}\{b(X^n) = 0\} = p$. Note that if k is even, b does not depend on its input bit x_n. Therefore, $T_\alpha(p)$ is well-defined for all dyadic rationals $p \in [0, 1]$. It is shown in Appendix D that $T_\alpha(\cdot)$ is continuous on the dyadic rationals (in fact, it is Hölder continuous with exponent 1/2). Therefore, $T_\alpha(p)$ is also well-defined when $p \in [0, 1]$ is not a dyadic rational, by considering its unique continuous extension to $[0, 1]$.

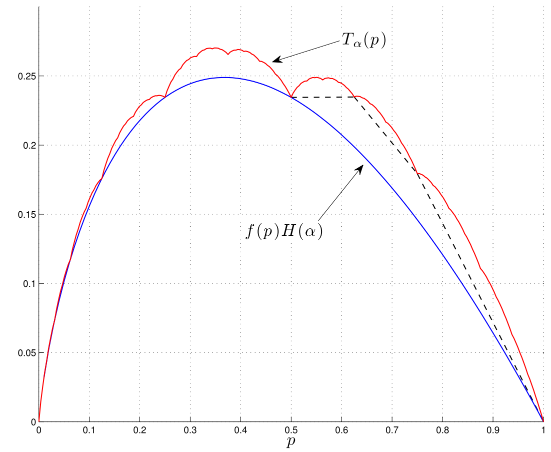

Fig. 2. A comparison of $T_\alpha(p)$ and $f(p)H(\alpha)$ for $\alpha = 0.1$. The broken line shows the three chords Algorithm II.1 constructs before terminating.

Thus, the validity of (18) for all lex b (and all n) is equivalent to the inequality

$$T_\alpha(p) \geq f(p) H(\alpha) \quad \forall p \in [0, 1], \tag{20}$$

motivating the following theorem.

Theorem 4: Fix $\alpha \in (0, 1/2)$. If a call to Algorithm II.1 with arguments $(p_-, p_+) = (1/2, 1)$ eventually terminates, then Conjecture 3 is true for the chosen α.

```
Algorithm II.1: TESTINEQUALITY($p_-$, $p_+$)

main
  if CHECKCHORD($p_-$, $p_+$) < 0
    then { $p \leftarrow \frac{1}{2}(p_- + p_+)$
         { TESTINEQUALITY($p_-$, $p$)
         { TESTINEQUALITY($p$, $p_+$)

procedure CHECKCHORD($a$, $b$)
  comment: $C(x)$ is the chord connecting
           the points $(a, T_\alpha(a))$ and $(b, T_\alpha(b))$.
  $C(x) := \frac{T_\alpha(b) - T_\alpha(a)}{b - a}(x - a) + T_\alpha(a)$
  $\nu \leftarrow \min_{x \in [a,b]} C(x) - f(x)H(\alpha)$
  return ($\nu$)
```

The key idea behind Theorem 4 is that Algorithm II.1 recursively constructs a piecewise linear function on the interval $p \in [1/2, 1]$ which simultaneously upper bounds $f(p)H(\alpha)$ and lower bounds $T_\alpha(p)$. As discussed in Section III-C, this is sufficient to prove (20) due to restricted-concavity and self-similarity characteristics of $T_\alpha(p)$. Figure 2 illustrates the function constructed by Algorithm II.1 for $\alpha = 0.1$.

[2] A threshold function is a Hamming ball (of any radius) centered at the all-zero vector. The radius determines the probability that the threshold function returns 0.

Using a Matlab implementation of Algorithm II.1, we have validated (20) for α ranging from 0 to 1/2 in increments of 0.001. Hence, it is reasonable to believe that Conjecture 3 is true in general.

Despite the apparent gap between $T_\alpha(p)$ and $f(p)H(\alpha)$ for $p \in (1/2, 1)$ (e.g., Fig. 2), the oscillatory behavior of $T_\alpha(p)$ seems to render traditional analysis techniques ineffective in establishing (20). This was our motivation for pursuing a computer-assisted method for verifying (7). To get a sense for the strange behavior of $T_\alpha(p)$, we show in Appendix B that $\lim_{\alpha\to 0} T_\alpha(p)/H(\alpha)$ is equal to the *Takagi function*, a classical construction of an everywhere-continuous, nowhere-differentiable function closely related to the edge-isoperimetric inequality given in Theorem 2 (see [12], [13]). Even for $\alpha = 0.1$, the pathological nature of $T_\alpha(p)$ makes itself apparent in Figure 2.

Remark 3: In the subroutine CHECKCHORD*(a, b) of Algorithm II.1, the minimization can be computed in closed form.*

Remark 4: Inherently, the running time of Algorithm II.1 will depend on α, but does not depend on n. Indeed, Theorem 4 states that termination of TESTINEQUALITY$\left(0, \frac{1}{2}\right)$ *guarantees validity of* (18) *for all $n = 1, 2, 3, \ldots$.*

III. PROOFS

In this section, we prove our main results.

A. Proof of Theorem 1

Here, we use Fourier-analytic techniques to prove Theorem 1, which we restate more precisely as the following proposition:

Proposition 1: Let X^n be i.i.d. Bernoulli(1/2), and let Y^n be the result of passing X^n through a memoryless binary symmetric channel with crossover probability α. For any Boolean function $b : \{0, 1\}^n \to \{0, 1\}$ with $\mathscr{P}\{b(X^n) = 0\} = \frac{1}{2}$, we have

$$\sum_{i=1}^{n} I(b(X^n); Y_i) \leq 1 - H(\alpha). \tag{21}$$

We will require the following lemma.

Lemma 1: Let $r^2 \leq 1$ and $\rho \in [0, 1]$. An optimal solution for the (non-convex) program

$$\text{minimize :} \quad \sum_{i=1}^{n} H\left(\frac{1+\rho x_i}{2}\right) \tag{22}$$

$$\text{subject to :} \quad \sum_{i=1}^{n} x_i^2 \leq r^2 \tag{23}$$

is given by $x_1 = r$, and $x_i = 0$ for $i = 2, \ldots, n$.

Proof: By considering the Taylor series expansion of the binary entropy function around 1/2, the objective function can be rewritten as

$$\sum_{i=1}^{n} H\left(\frac{1+\rho x_i}{2}\right) = n - \log e \sum_{k=1}^{\infty} \sum_{i=1}^{n} \frac{(\rho x_i)^{2k}}{(2k-1)2k}. \tag{24}$$

For each fixed k, the sum $\sum_{i=1}^{n} x_i^{2k}$ is convex in the squared x_i's, and is therefore maximized subject to the constraint $\sum_{i=1}^{n} x_i^2 \leq r^2$ when $x_1 = r$, and $x_i = 0$ for $i = 2, \ldots, n$. □

We are now in a position to prove Theorem 1.

Proof of Theorem 1: We prove the claim using the Fourier transform. For convenience, we will assume X^n and Y^n take values in $\{-1, 1\}^n$. Consequently, we consider Boolean functions $b : \{-1, 1\}^n \to \{-1, 1\}$. Any Boolean function $b : \{-1, 1\}^n \to \{-1, 1\}$ can be written in terms of its Fourier coefficients as (see [8])

$$b(x^n) = \sum_{S\subseteq[1:n]} \hat{b}(S)\Pi_S(x^n),$$

where $\Pi_S(x^n) = \prod_{i\in S} x_i$ are the orthonormal basis functions for the Fourier transform, and $\{\hat{b}(S)\}_{S\subseteq[1:n]}$ are the Fourier coefficients defined by

$$\hat{b}(S) = \mathbb{E}b(X^n)\Pi_S(X^n). \tag{25}$$

If $S = \emptyset$, then we define $\Pi_S(x^n) = 1$. It is easy to verify the following identities, the first two of which verify that $\Pi_S(x^n) = \prod_{i\in S} x_i$ are indeed orthonormal basis functions:

1) For all x^n, $\Pi_S(x^n)^2 = 1$ since $\Pi_S(x^n) \in \{-1, 1\}$. Therefore,

$$\mathbb{E}\Pi_S(X^n)^2 = 1. \tag{26}$$

2) If $S \neq T$ and neither are equal to $\emptyset$, then $\Pi_S(X^n)$ and $\Pi_T(X^n)$ are independent and uniformly distributed on $\{-1, 1\}$ giving the identity

$$\mathbb{E}\Pi_S(X^n)\Pi_T(X^n) = 0. \tag{27}$$

On the other hand, if $\emptyset = S \neq T$, then we simply have $\mathbb{E}\Pi_S(X^n)\Pi_T(X^n) = \mathbb{E}\Pi_T(X^n) = 0$.

3) As a consequence, Parseval's theorem holds:

$$1 = \mathbb{E}b(X^n)^2 = \textstyle\sum_{S\subseteq[1:n]} \hat{b}(S)^2, \tag{28}$$

since

$$\begin{aligned}
&\mathbb{E}\big[b(X^n)^2\big]\\
&\quad= \mathbb{E}\left[\left(\sum_{S\subseteq[1:n]} \hat{b}(S)\Pi_S(X^n)\right)\left(\sum_{T\subseteq[1:n]} \hat{b}(T)\Pi_T(X^n)\right)\right]\\
&\quad= \sum_{S\subseteq[1:n]}\sum_{T\subseteq[1:n]} \hat{b}(S)\hat{b}(T)\mathbb{E}\left[\Pi_S(X^n)\Pi_T(X^n)\right] \qquad (29)\\
&\quad= \sum_{S\subseteq[1:n]} \hat{b}(S)^2. \qquad (30)
\end{aligned}$$

By definition of the Fourier coefficient $\hat{b}(\emptyset)$, we have $\hat{b}(\emptyset) = \mathbb{E}\left[b(X^n)\right] = 0$ since $b(X^n)$ is equiprobable on $\{-1, 1\}$. Also,

$$\begin{aligned}
&\mathscr{P}\{b(X^n) = 1|Y_i = y_i\}\\
&\quad= \frac{1 + \mathbb{E}\left[b(X^n)|Y_i = y_i\right]}{2} \qquad (31)\\
&\quad= \frac{1 + \mathbb{E}\left[\sum_{S\subseteq[1:n]} \hat{b}(S)\Pi_S(X^n)\Big|Y_i = y_i\right]}{2} \qquad (32)\\
&\quad= \frac{1 + \sum_{S\subseteq[1:n]} \hat{b}(S)\mathbb{E}\left[\Pi_S(X^n)\Big|Y_i = y_i\right]}{2} \qquad (33)
\end{aligned}$$

Observe that $\mathbb{E}\left[\Pi_S(X^n)|Y_i = y_i\right] = 0$ unless $S = \emptyset$ or $S = \{i\}$ since $\{X_j\}_{j\in S\setminus\{i\}}$ are independent from Y_i. Hence, we have

$$\mathscr{P}\{b(X^n) = 1|Y_i = y_i\} = \frac{1 + \hat{b}(\emptyset) + \rho y_i \hat{b}(\{i\})}{2} \quad (34)$$

$$= \frac{1 + \rho y_i \hat{b}(\{i\})}{2}, \quad (35)$$

where $\rho \triangleq \mathbb{E}[X_i|Y_i = 1] = 1 - 2\alpha$. By definition of mutual information and our assumption that $\mathscr{P}\{b(X^n) = 1\} = \frac{1}{2}$,

$$\sum_{i=1}^{n} I(b(X^n); Y_i) = nH(b(X^n)) - \sum_{i=1}^{n} H(b(X^n)|Y_i) \quad (36)$$

$$= n - \sum_{i=1}^{n} H(b(X^n)|Y_i). \quad (37)$$

Hence, we prove (21) by attempting to minimize $\sum_{i=1}^{n} H(b(X^n)|Y_i)$ for $\mathscr{P}\{b(X^n) = 1\} = \frac{1}{2}$. Using (35), observe that

$$\sum_{i=1}^{n} H(b(X^n)|Y_i) \quad (38)$$

$$= \sum_{i=1}^{n} \frac{1}{2}\Big[H\left(\mathscr{P}\{b(X^n) = 1|Y_i = 1\}\right) \quad (39)$$

$$+ H\left(\mathscr{P}\{b(X^n) = 1|Y_i = -1\}\right)\Big]$$

$$= \sum_{i=1}^{n} \frac{1}{2}\left[H\left(\frac{1 + \rho\hat{b}(\{i\})}{2}\right) + H\left(\frac{1 - \rho\hat{b}(\{i\})}{2}\right)\right] \quad (40)$$

$$= \sum_{i=1}^{n} H\left(\frac{1 + \rho\hat{b}(\{i\})}{2}\right), \quad (41)$$

where the last equality holds since

$$1 - \left(\frac{1 - \rho\hat{b}(\{i\})}{2}\right) = \left(\frac{1 + \rho\hat{b}(\{i\})}{2}\right). \quad (42)$$

Recalling that Parseval's identity implies $\sum_{i=1}^{n} \hat{b}(\{i\})^2 \le 1$, an application of Lemma 1 implies that (41) is minimized by setting $\hat{b}(\{1\}) = 1$ and $\hat{b}(\{i\}) = 0$ for $i = 2, \ldots, n$. Thus,

$$\sum_{i=1}^{n} H(b(X^n)|Y_i) \ge H(\alpha) + (n - 1). \quad (43)$$

Substitution into (37) proves the theorem. □

B. Proof of Theorem 3

We begin the proof of Theorem 3 by first proving the following result for 1-dimensional compressions. We simplify notation by writing $\mathscr{P}\{E|y^n\}$ to denote the probability of the event E conditioned on $Y^n = y^n$.

Lemma 2: Let $b : \Omega_n \to \Omega$ and $i \in \{1, 2, \ldots, n\}$. If $\hat{b} : \Omega_n \to \Omega$ is defined by its preimage $\hat{b}^{-1}(0) = C_{\{i\}}(b^{-1}(0))$, then $I(\hat{b}(X^n); Y^n) \ge I(b(X^n); Y^n)$.

Proof: It suffices to consider the case where $i = n$, as any other case can be handled by first permuting coordinates.

Define $B = b^{-1}(0)$ and $\hat{b}^{-1}(0) = C_{\{n\}}(B)$, and let

$$E_0 = \left\{x^{n-1} : B_{\{n\}}(x^{n-1}, 0) = \{0\}\right\} \quad (44)$$

$$E_1 = \left\{x^{n-1} : B_{\{n\}}(x^{n-1}, 0) = \{1\}\right\}, \quad (45)$$

where $B_{\{n\}}$ is defined by (8) with $\mathcal{I} = \{n\}$. Define $\breve{b}(x^{n-1}, x_n) := \hat{b}(x^{n-1}, \neg x_n)$, where $\neg x_n$ is the negation of x_n.

By definition of $\hat{b}$ and $\breve{b}$, we have the identities

$$\mathscr{P}\{b(X^n) = 0|y^{n-1}, 0\}$$

$$= \mathscr{P}\{\hat{b}(X^n) = 0|y^{n-1}, 0\} - (1 - 2\alpha)\mathscr{P}\{X^{n-1} \in E_1|y^{n-1}\} \quad (46)$$

$$= \mathscr{P}\{\breve{b}(X^n) = 0|y^{n-1}, 0\} + (1 - 2\alpha)\mathscr{P}\{X^{n-1} \in E_0|y^{n-1}\}. \quad (47)$$

The identity (46) follows since

$$\begin{aligned}
\mathscr{P}\{\hat{b}(X^n) = 0|y^{n-1}, 0\}& \\
= \mathscr{P}\{b(X^n) = 0|y^{n-1}, 0\}& \quad (48)\\
+ \mathscr{P}\{X_n = 0|Y_n = 0\}\mathscr{P}\{X^{n-1} \in E_1|y^{n-1}\}& \\
- \mathscr{P}\{X_n = 1|Y_n = 0\}\mathscr{P}\{X^{n-1} \in E_1|y^{n-1}\}& \\
= \mathscr{P}\{b(X^n) = 0|y^{n-1}, 0\}& \\
+ (1 - 2\alpha)\mathscr{P}\{X^{n-1} \in E_1|y^{n-1}\}&
\end{aligned}$$

by construction. Identity (47) follows similarly.

Similar identities hold for $\mathscr{P}\{b(X^n) = 0|y^{n-1}, 1\}$ with opposite signs on the $(1 - 2\alpha)$ terms, giving

$$\mathscr{P}\{b(X^n) = 0|y^n\}$$

$$= \theta\mathscr{P}\{\hat{b}(X^n) = 0|y^n\} + (1 - \theta)\mathscr{P}\{\breve{b}(X^n) = 0|y^n\}, \quad (49)$$

where

$$\theta = \frac{\mathscr{P}\{X^{n-1} \in E_0|y^{n-1}\}}{\mathscr{P}\{X^{n-1} \in E_0|y^{n-1}\} + \mathscr{P}\{X^{n-1} \in E_1|y^{n-1}\}}. \quad (50)$$

Concavity of entropy implies that

$$\theta H(\hat{b}(X^n)|y^n) + (1 - \theta)H(\breve{b}(X^n)|y^n) \le H(b(X^n)|y^n).$$

Noting that θ only depends on y^{n-1}, we average both sides over $y_n \in \{0, 1\}$ to obtain

$$\theta H(\hat{b}(X^n)|y^{n-1}, Y_n) + (1 - \theta)H(\breve{b}(X^n)|y^{n-1}, Y_n)$$

$$\le H(b(X^n)|y^{n-1}, Y_n). \quad (51)$$

By symmetry, $H(\hat{b}(X^n)|y^{n-1}, Y_n) = H(\breve{b}(X^n)|y^{n-1}, Y_n)$. Therefore, averaging (51) over all values of y^{n-1}, we can conclude that $H(\hat{b}(X^n)|Y^n) \le H(b(X^n)|Y^n)$. To complete the proof, we recall that $|\hat{b}^{-1}(0)| = |C_{\{n\}}(b^{-1}(0))| = |b^{-1}(0)|$. Combined with the fact that X^n is uniformly distributed on Ω_n, this implies that $H(\hat{b}(X^n)) = H(b(X^n))$, as desired. □

We are now in a position to finish the proof of Theorem 3, which is similar to the proof of Lemma 2.

Proof of Theorem 3: We assume that $\mathcal{I} = \{n - 1, n\}$, as all other cases follow by a permutation of coordinates. To simplify notation, we write $B = b^{-1}(0)$.

By a repeated application of Lemma 2, we can assume that B is $\{n - 1\}$- and $\{n\}$-compressed. Indeed, if $\{i\}$-compression changes an element of B, it moves it *lower* in the lexicographical ordering. Thus, if we iteratively apply $\{n - 1\}$- and $\{n\}$-compressions to B, we must eventually reach a fixed point.

Thus, the $\mathcal{I}$-sections $B_{\mathcal{I}}(x^n)$ can only be one of the following: $\emptyset$, $\{00\}$, $\{00, 01\}$, $\{00, 10\}$, $\{00, 01, 10\}$, or $\{00, 01, 10, 11\}$. Note that all of these sets are initial segments of the lexicographical order on Ω_2 except $\{00, 10\}$. Hence, we aim to transform B so that $B_{\mathcal{I}}(x^n) \neq \{00, 10\}$. To this end, define

$$G_0 = \left\{x^n : x_{n-1} = x_n = 0, B_{\mathcal{I}}(x^n) = \{00, 01\}\right\} \quad (52)$$
$$G_1 = \left\{x^n : x_{n-1} = x_n = 0, B_{\mathcal{I}}(x^n) = \{00, 10\}\right\}. \quad (53)$$

Now, define $\hat{b}$ by $\hat{b}^{-1}(0) = C_{\mathcal{I}}(B)$ and the function $\breve{b}$ by permuting the last two coordinates:

$$\breve{b}(x^{n-2}, x_n, x_{n-1}) = \hat{b}(x^{n-2}, x_{n-1}, x_n). \quad (54)$$

By an argument similar to that given in (46)-(50), it follows that

$$\mathscr{P}\{b(X^n) = 0|y^n\} = \theta\mathscr{P}\{\hat{b}(X^n) = 0|y^n\} + (1-\theta)\mathscr{P}\{\breve{b}(X^n) = 0|y^n\}, \quad (55)$$

where

$$\theta = \frac{\mathscr{P}\{X^{n-2} \in G_0|y^{n-2}\}}{\mathscr{P}\{X^{n-2} \in G_0|y^{n-2}\} + \mathscr{P}\{X^{n-2} \in G_1|y^{n-2}\}}. \quad (56)$$

Concavity of entropy implies that

$$\theta H(\hat{b}(X^n)|y^n) + (1-\theta)H(\breve{b}(X^n)|y^n) \leq H(b(X^n)|y^n).$$

Noting that θ only depends on y^{n-2}, we average both sides over y_{n-1}, y_n to obtain

$$\theta H(\hat{b}(X^n)|y^{n-2}, Y_{n-1}^n) + (1-\theta)H(\breve{b}(X^n)|y^{n-2}, Y_{n-1}^n) \leq H(b(X^n)|y^{n-2}, Y_{n-1}^n). \quad (57)$$

Crucially, the symmetry (54) implies that

$$H(\hat{b}(X^n)|y^{n-2}, Y_{n-1}^n) = H(\breve{b}(X^n)|y^{n-2}, Y_{n-1}^n). \quad (58)$$

Combining this with (57) and averaging over y^{n-2} proves $H(\hat{b}(X^n)|Y^n) \leq H(b(X^n)|Y^n)$. Since $H(\hat{b}(X^n)) = H(b(X^n))$, the proof is complete. □

C. Proof of Theorem 4

The proof of Theorem 4 requires the following lemmas.

Lemma 3: The following identity holds:

$$\tfrac{1}{2}\Big[T_\alpha(2p) - f(2p)H(\alpha)\Big] = \Big[T_\alpha(p) - f(p)H(\alpha)\Big]. \quad (59)$$

Proof: Suppose b and b' are both lex and satisfy[3]

$$p \triangleq \mathscr{P}\{b'(X^n) = 0\} = \frac{1}{2}\mathscr{P}\{b(X^{n-1}) = 0\}. \quad (60)$$

We have the identities $f(pq) = pf(q) + qf(p)$ and

$$\mathscr{P}\{b'(X^n) = 0|y^n\} = \mathscr{P}\{X_1 = 0|y_1\}\mathscr{P}\{b(X_2^n) = 0|y_2^n\},$$

which imply the relation: $2T_\alpha(p) = T_\alpha(2p) + 2pH(\alpha)$. The claim now follows easily. □

Although $T_\alpha(p)$ is not concave, we are able to prove a restricted-concavity characteristic of $T_\alpha(p)$. This is exploited in the following claim.

Lemma 4: Let $k < 2^n$, and define $p_- \triangleq k2^{-n}$ and $p_+ \triangleq (k+1)2^{-n}$. For $\theta \in [0, 1]$, the following inequality holds:

$$T_\alpha(\theta p_- + (1-\theta)p_+) \geq \theta T_\alpha(p_-) + (1-\theta)T_\alpha(p_+). \quad (61)$$

Proof: First, observe that it suffices to prove

$$T_\alpha\left(\frac{p_- + p_+}{2}\right) \geq \frac{1}{2}\Big[T_\alpha(p_-) + T_\alpha(p_+)\Big]. \quad (62)$$

Indeed, from (62) an inductive argument proves the lemma when $\theta p_- + (1-\theta)p_+$ is restricted to the set of dyadic rationals. Then, recalling the continuity of $T_\alpha(\cdot)$ on $[0, 1]$ completes the proof (see Appendix D).

To this end, let b be the unique lex function on $n+1$ inputs which satisfies

$$\mathscr{P}\{b(X^{n+1}) = 0\} = \frac{2k+1}{2^{n+1}} = \frac{1}{2}\Big[p_- + p_+\Big], \quad (63)$$

and let b_-, b_+ be the unique lex functions on n inputs which satisfy

$$\mathscr{P}\{b_-(X^n) = 0\} = p_- \quad \mathscr{P}\{b_+(X^n) = 0\} = p_+. \quad (64)$$

By construction, we have

$$\begin{aligned}\mathscr{P}\{b(X^{n+1}) &= 0|Y^{n+1}\} \\ &= \mathscr{P}\{X_{n+1} = 0|Y_{n+1}\}\mathscr{P}\{b_+(X_1^n) = 0|Y_1^n\} \\ &\quad + \mathscr{P}\{X_{n+1} = 1|Y_{n+1}\}\mathscr{P}\{b_-(X_1^n) = 0|Y_1^n\}.\end{aligned} \quad (65)$$

Combining (63) and (65) with the fact that $f(x)$ is concave, we have the desired inequality

$$\begin{aligned}T_\alpha&\left(\frac{p_- + p_+}{2}\right) \\ &\geq \mathbb{E}_{Y^{n+1}}\mathscr{P}\{X_{n+1} = 0|Y_{n+1}\}f\left(\mathscr{P}\{b_+(X_1^n) = 0|Y_1^n\}\right) \\ &\quad + \mathbb{E}_{Y^{n+1}}\mathscr{P}\{X_{n+1} = 1|Y_{n+1}\}f\left(\mathscr{P}\{b_-(X_1^n) = 0|Y_1^n\}\right) \\ &= \frac{1}{2}\Big[T_\alpha(p_-) + T_\alpha(p_+)\Big].\end{aligned} \quad (66)$$

□

We are now in a position to prove Theorem 4.

Proof of Theorem 4: Lemma 3 implies that if (20) holds for p, then it also holds for $2p$, and hence it suffices to consider $p \in [1/2, 1]$. Therefore, Lemma 4 implies that it is sufficient to construct a piecewise linear function $g : [1/2, 1] \to [0, \infty)$ satisfying the following properties:

1) Each segment of g is a chord connecting the points $(p_-, T_\alpha(p_-))$ and $(p_+, T_\alpha(p_+))$, where p_- and p_+ are of the form

$$p_- = \frac{k}{2^n}, \qquad p_+ = \frac{k+1}{2^n}. \quad (67)$$

for integers k, n satisfying $k < 2^n$.

2) For $p \in [1/2, 1]$, $g(p) \geq f(p)H(\alpha)$.

By definition, Algorithm II.1 terminates only if it constructs such a function. □

[3] Any lex function with $\mathscr{P}\{b(X^n) = 0\} = k/2^n$ can be reduced to a lex function on $n-1$ inputs if k is even.

IV. Concluding Remarks

Although Conjecture 1 remains open, we have provided evidence in support of its validity. Indeed, our results allow us to exhaustively verify Conjecture 2 for $n \leq 7$, and we have a computer-assisted method for establishing Conjecture 3 for any given value of α. Any complete proof of Conjectures 1 or 2 would be of significant interest, since it would likely require new methods which may be applicable in information theory and elsewhere (e.g., in proving discrete isoperimetric inequalities). An analytical proof of Conjecture 3 which does not require computer assistance would also be interesting, but appears difficult given the pathological nature of $H(b(X^n)|Y^n)$ for lex functions.

In closing, we propose two weaker forms of Conjecture 1 which could provide insight, but are still open.

1) Does Theorem 1 continue to hold when $b(X^n)$ is not equiprobable? Unfortunately, our Fourier-analytic proof of Theorem 1 appears to fail in this setting. Nonetheless, we feel that establishing this generalization of Theorem 1 should be considerably easier than establishing the stronger statement of Conjecture 1.

2) For Boolean functions b_1, b_2, does it hold that

$$I(b_1(X^n); b_2(Y^n)) \leq 1 - H(\alpha)? \tag{68}$$

While this problem appears difficult in general, it is a simple exercise to show this is true when $b_1(X^n)$ and $b_2(Y^n)$ are both equiprobable by using the data-processing property of maximal correlation and the fact that the only couplings between $b_1(X^n)$ and $b_2(Y^n)$ are symmetric. Intuitively, this should be the case for b_1, b_2 which maximize $I(b_1(X^n); b_2(Y^n))$, but a proof remains elusive.

Recent Progress

While this paper was under review, there were several related developments. We summarize them below.

In the time since (68) was first suggested by the authors in [1], Anantharam, Gohari, Kamath and Nair have made some tangible progress. In particular, they leveraged hypercontractivity to numerically verify (68), further supporting the conjectures in this paper. Interested readers are referred to their recent paper [14] for details.

Separately, Bogdanov and Nair have proven (68) holds in the restrictive setting of $b_1 = b_2$ [15] and $\alpha \leq 1/2$ using a Fourier-analytic argument. In a correspondence with C. Nair, the first author observed that Bogdanov and Nair's argument readily generalizes to the setting where b_1, b_2 are Boolean functions satisfying $\mathscr{P}\{b_1(X^n) = 0\} = \mathscr{P}\{b_2(X^n) = 0\}$, provided that the correlation inequality

$$\begin{aligned}\mathscr{P}\{b_1(X^n) &= b_2(X^n) = 0\} \\ &\geq \mathscr{P}\{b_1(X^n) = 0\}\mathscr{P}\{b_2(X^n) = 0\}\end{aligned} \tag{69}$$

holds. This latter result was proved independently by Calmon, Varia, and Médard in a recent preprint [16].

On this note, we remark that it suffices to consider monotone Boolean functions b_1, b_2 (one increasing and the other decreasing) in (68) when (69) does not hold. This is easily established using a rearrangement argument similar to our proof of Lemma 2, and may help fully resolve the setting where $\mathscr{P}\{b_1(X^n) = 0\} = \mathscr{P}\{b_2(X^n) = 0\}$.

Finally, it was claimed in a presentation by Chandar and Tchamkerten [17] that $H(b(X^n)|Y^n)$ is not necessarily minimized by lex functions when $\mathscr{P}\{b(X^n) = 0\}$ is sufficiently close to zero. The definition of “sufficiently close” depended on α, and their argument involved taking $\mathscr{P}\{b(X^n) = 0\}$ exponentially small in n. As such, Chandar and Tchamkerten's work does not address the setting in Conjecture 2 where

$$H(b(X^n)) \geq 1 - H(\alpha), \tag{70}$$

implying that it remains open at the time of press. Since it suffices to consider functions b satisfying (70) when addressing Conjecture 1, the claim of Chandar and Tchamkerten does not immediately apply to this conjecture either.

In any case, if Conjecture 2 is discovered to be false, it may be true that $H(b(X^n)|Y^n)$ is always minimized by a function that is “almost lex" (i.e., in the sense that $b^{-1}(0)$ is a subset of the hypercube with a small edge-boundary [18]). This could still lead to a proof of Conjecture 1, which was our initial motivation.

Appendix A
Randomized Boolean functions do Not Improve Mutual Information

As mentioned briefly in the introduction, it suffices to consider deterministic functions when upper-bounding the quantity $I(b(X^n); Y^n)$. Indeed, any randomized Boolean function b can be written as $\tilde{b}(Q, X^n)$, where $\tilde{b}$ is a deterministic function and Q is a random variable that is independent of X^n. In this case, we have

$$\begin{aligned} I(\tilde{b}(Q, X^n); Y^n) &\leq I(\tilde{b}(Q, X^n), Q; Y^n) && (71)\\ &= I(Q; Y^n) + I(\tilde{b}(Q, X^n); Y^n|Q) && (72)\\ &= I(\tilde{b}(Q, X^n); Y^n|Q). && (73)\end{aligned}$$

Thus, there must exist some q such that

$$I(\tilde{b}(Q, X^n); Y^n|Q = q) \geq I(\tilde{b}(Q, X^n); Y^n), \tag{74}$$

which establishes the claim that randomization does not help.

Appendix B
Further Comments on Edge-Isoperimetry, the Takagi Function, and $H(b(X^n)|Y^n)$

The Takagi function is a simple construction of an everywhere continuous, nowhere differentiable function [19]. Specifically, the Takagi function $\mathbb{T}(p)$ is defined on the unit interval [0, 1] by

$$\mathbb{T}(p) = \sum_{n=0}^{\infty} \frac{s(p2^n)}{2^n}, \tag{75}$$

where $s(x) = \min_{m \in \mathbb{Z}} |m - x|$ (i.e., the distance to the nearest integer). In 2000, Guu [13] connected the Takagi function with the edge-isoperimetry problem mentioned in Section II-C.[4] We recall his result in the following theorem.

Theorem 5 (From [13]):

$$\min_{S\subseteq V(Q_n):|S|=k} \frac{|\partial(S)|}{2^n} = \mathbb{T}\left(\frac{k}{2^n}\right) \tag{76}$$

As we alluded to in Section II-C, $H(b(X^n)|Y^n)$ is related to the edge-isoperimetric problem and, consequently, the Takagi function. The relationship becomes most apparent in the limiting case of $\alpha \to 0$, which is made precise by the following theorem.

Theorem 6: For any Boolean function b, associate $b^{-1}(0)$ with the corresponding set of vertices in the n-dimensional hypercube Q_n. Let $|\partial(b^{-1}(0))|$ be the size of the edge boundary of $b^{-1}(0)$ in Q_n. Then

$$\begin{aligned}\lim_{\alpha\to 0} \frac{H(b(X^n)|Y^n)}{H(\alpha)} &= 2\frac{|\partial(b^{-1}(0))|}{2^n} && (77)\\ &\geq 2\mathbb{T}\left(\frac{|b^{-1}(0)|}{2^n}\right) && (78)\\ &= 2\mathbb{T}\Big(\mathscr{P}\left\{b(X^n)=0\right\}\Big). && (79)\end{aligned}$$

Moreover, the lower bound is attained with equality if b is lex.

In fact, it is known that $|\partial(S)|2^{-n} = \mathbb{T}\left(|S|2^{-n}\right)$ only if S coincides with an initial segment of the lexicographical ordering under an isomorphism of the hypercube Q_n. Therefore, we can infer from Theorem 6 that Conjecture 2 is valid for sufficiently small α since those functions b which are not equivalent to a lex function (in the same sense as above) will not meet the lower bound (79) as α vanishes.

Remark 5: Before continuing to the proof, we remark that information-theoretic techniques have been applied to proving edge-isoperimetric inequalities in the past [21], [22]. However, it appears that those results are related to our problem only insofar as Harper's edge-isoperimetric inequality is related.

Theorem 6 follows immediately from Theorem 2, Theorem 5, and the following lemma.

Lemma 5:

$$\lim_{\alpha\to 0} \frac{\mathbb{E}_{Y^n}[f(\mathscr{P}\{b(X^n)=0|Y^n\})]}{H(\alpha)} = \frac{|\partial(b^{-1}(0))|}{2^n}, \tag{80}$$

where $f(x) = x\log(1/x)$.

Proof: For convenience let $\bar{\alpha} = 1-\alpha$. By definition, we know that

$$\mathscr{P}(b(X^n)=0|y^n) = \sum_{x^n\in b^{-1}(0)} \alpha^{d(x^n,y^n)}\bar{\alpha}^{n-d(x^n,y^n)}, \tag{81}$$

where $d(x^n, y^n)$ is the Hamming distance between vectors x^n and y^n. Suppose $\alpha^k\bar{\alpha}^{n-k}$ is the largest term in the sum (81) for a given y^n, and put $N_k(y^n) = \left|\{x^n \in b^{-1}(0) : d(x^n,y^n)=k\}\right|$. Then, for α near zero and $k \geq 1$,

$$\begin{aligned}&f(\mathscr{P}(b(X^n)=0|y^n)) && (82)\\ &= f\left(N_k(y^n)\alpha^k\bar{\alpha}^{n-k} + O(\alpha^{k+1})\right) && (83)\\ &= -\left(N_k(y^n)\alpha^k\bar{\alpha}^{n-k} + O(\alpha^{k+1})\right) && (84)\\ &\quad\times\Big[\log\left(N_k(y^n)\alpha^k\bar{\alpha}^{n-k}\right) + \log\left(1+O(\alpha)\right)\Big]\\ &= -\left(N_k(y^n)\alpha^k + O(\alpha^{k+1})\right) && (85)\\ &\quad\times\Big[k\log(\alpha) + \log(N_k(y^n)) + O(\alpha)\Big]\\ &= -kN_k(y^n)\alpha^k\log(\alpha) + O(\alpha^k) + O\left(\alpha^{k+1}\log(\alpha)\right). && (86)\end{aligned}$$

Therefore,

$$\begin{aligned}&\frac{f(\mathscr{P}(b(X^n)=0|y^n))}{H(\alpha)}\\ &\quad= \frac{-kN_k(y^n)\alpha^k\log(\alpha) + O(\alpha^k) + O\left(\alpha^{k+1}\log(\alpha)\right)}{-\alpha\log(\alpha) + o(\alpha\log(\alpha))} && (87)\\ &\quad\xrightarrow{\alpha\to 0} \begin{cases} N_1(y^n) & \text{if } k=1\\ 0 & \text{if } k>1. \end{cases} && (88)\end{aligned}$$

On the other hand, if $k=0$, then $y^n \in b^{-1}(0)$, and $N_0(y^n)=1$. In this case,

$$\begin{aligned}f(\mathscr{P}(b(X^n)=0|y^n)) &= f(\bar{\alpha}^n + O(\alpha)) && (89)\\ &= f(1+O(\alpha)) && (90)\\ &= -(1+O(\alpha))\log(1+O(\alpha)) && (91)\\ &= O(\alpha) && (92)\\ &\xrightarrow{\alpha\to 0} 0. && (93)\end{aligned}$$

Since $\mathscr{P}\{Y^n = y^n\} = 2^{-n}$ for all $y^n \in \{0,1\}^n$, we can conclude that

$$\lim_{\alpha\to 0} \frac{\mathbb{E}_{Y^n}[f(\mathscr{P}\{b(X^n)=0|Y^n\})]}{H(\alpha)} = \frac{|\partial(b^{-1}(0))|}{2^n} \tag{94}$$

as claimed. □

Remark 6: Using a Taylor series expansion similar to the proof above, it is possible to prove that Conjecture 1 holds as $\alpha \to 1/2$. However, since this setting does not exhibit the nice connection to the edge-isoperimetric inequality that occurs when $\alpha \to 0$, we omit the details.

Appendix C
Enumerating the Functions in $\mathcal{S}_n$

Recall from Corollary 1 that $\mathcal{S}_n$ denotes the set of Boolean functions on n inputs for which $b^{-1}(0)$ is $\mathcal{I}$-compressed for all $\mathcal{I}$ of cardinality 2. Suppose $b : \Omega_{n+1} \to \Omega$, and define the functions $b_0 : \Omega_n \to \Omega$ and $b_1 : \Omega_n \to \Omega$ by

$$\begin{aligned}b_0(x^n) &= b(0,x^n) && (95)\\ b_1(x^n) &= b(1,x^n) && (96)\end{aligned}$$

for all $x^n \in \Omega_n$. By definition, $b \in \mathcal{S}_{n+1}$ if, and only if, $b_0 \in \mathcal{S}_n$, $b_1 \in \mathcal{S}_n$, and $b^{-1}(0)$ is $\{1,j\}$-compressed for all $j \in \{2,3,\ldots,n+1\}$. Therefore, given $\mathcal{S}_n$, we can enumerate all functions in $\mathcal{S}_{n+1}$ in $O(n\,2^{n-1}\,|\mathcal{S}_n|^2)$ time since,

[4] Takagi-like functions also appear in various other problems (see [20]).

for $b : \Omega_{n+1} \to \Omega$, we can naïvely test whether $b^{-1}(0)$ is $\{1, j\}$-compressed in $O(2^{n-1})$ time. With this observation in mind, it is a routine exercise to recursively enumerate the functions in $\mathcal{S}_n$ for modest values of n.

Appendix D
Hölder Continuity of $T_\alpha(p)$

Let $f(x) = x\log(1/x)$. Suppose $p \in [0, 1]$ is a dyadic rational (i.e., $p = k2^{-n}$ for some nonnegative integers k, n). Recall our definition $T_\alpha(p) \triangleq \mathbb{E}_{Y^n} f(\mathscr{P}(b(X^n) = 0|Y^n))$, where $b : \{0, 1\}^n \to \{0, 1\}$ is a lex function satisfying $\mathscr{P}\{b(X^n) = 0\} = p$. Note that if k is even, then b does not depend on its input bit x_n. Therefore, by induction, $T_\alpha(p)$ is well-defined for all dyadic rationals $p \in [0, 1]$. As we show shortly, $T_\alpha(p)$ is continuous, and therefore has a unique continuous extension to the interval $[0, 1]$. In order to do so, we require the following lemma:

Lemma 6: For $x, y \in [0, 1]$,

$$|x\log(x) - y\log(y)| \le 2\sqrt{|x-y|}. \tag{97}$$

Proof: Assume without loss of generality that $x < y$. It is convenient to express $y = x + \delta$, where $\delta > 0$. Now, note that

$$\frac{\partial}{\partial x}\left(x\log(x) - (x+\delta)\log(x+\delta)\right) = \log\left(\frac{x}{x+\delta}\right) < 0,$$

and hence $|x\log(x) - (x+\delta)\log(x+\delta)|$ attains it's maximum when $x = 0$ or $x = 1 - \delta$. Thus, we can conclude that

$$\begin{aligned}&|x\log(x) - (x+\delta)\log(x+\delta)|\\ &\quad\le \max\left\{\delta\log\frac{1}{\delta}, (1-\delta)\log\frac{1}{1-\delta}\right\} \le 2\sqrt{\delta}. \end{aligned}\tag{98}$$

□

Lemma 7: $T_\alpha(p)$ is Hölder continuous. Specifically,

$$|T_\alpha(x) - T_\alpha(y)| \le 2\sqrt{|x-y|}. \tag{99}$$

Proof: We will prove the claim when x, y are dyadic rationals. Since dyadic rationals are dense, Hölder continuity of the same order carries over immediately to the continuous extension of T_α on the interval $[0, 1]$. To this end, let x, y be dyadic rationals with $x > y$, and let b_x and b_y be lex functions on $\{0, 1\}^n$ which satisfy $x = \mathscr{P}\{b_x(X^n) = 0\}$ and $y = \mathscr{P}\{b_y(X^n) = 0\}$. Then

$$\begin{aligned}&|T_\alpha(x) - T_\alpha(y)|\\ &= \left|\mathbb{E}_{Y^n}\Big[f(\mathscr{P}\{b_x(X^n) = 0|Y^n\}) - f(\mathscr{P}\{b_y(X^n) = 0|Y^n\})\Big]\right|\\ &\le \mathbb{E}_{Y^n}\left|f(\mathscr{P}\{b_x(X^n) = 0|Y^n\}) - f(\mathscr{P}\{b_y(X^n) = 0|Y^n\})\right|\\ &\le \mathbb{E}_{Y^n} 2\sqrt{\mathscr{P}\{b_x(X^n) = 0|Y^n\} - \mathscr{P}\{b_y(X^n) = 0|Y^n\}}\\ &\le 2\sqrt{\mathbb{E}_{Y^n}\Big[\mathscr{P}\{b_x(X^n) = 0|Y^n\} - \mathscr{P}\{b_y(X^n) = 0|Y^n\}\Big]}\\ &= 2\sqrt{\mathscr{P}\{b_x(X^n) = 0\} - \mathscr{P}\{b_y(X^n) = 0\}}\\ &= 2\sqrt{|x-y|}.\end{aligned}$$

In the above,

- the first inequality from the triangle inequality,
- the second inequality follows from Lemma 6 and the fact that $\mathscr{P}\{b_x(X^n) = 0|Y^n\} \ge \mathscr{P}\{b_y(Y^n) = 0|Y^n\}$ since $x > y$ implies $b_y^{-1}(0) \subset b_x^{-1}(0)$, and
- the third inequality follows by concavity of $\sqrt{x}$ for $x \ge 0$.

□

Acknowledgment

The authors would like to thank Abbas El Gamal, Chandra Nair, Sudeep Kamath, and Yeow-Khiang Chia for stimulating discussions. The authors would also like to thank an anonymous reviewer who suggested the present proof of Lemma 1 and provided several other constructive comments that improved this paper.

References

[1] G. R. Kumar and T. A. Courtade, "Which Boolean functions are most informative?" in *Proc. IEEE ISIT*, Jul. 2013, pp. 226–230.

[2] J. G. Klotz, D. Kracht, M. Bossert, and S. Schober, "Canalizing boolean functions maximize mutual information," *IEEE Trans. Inf. Theory*, vol. 60, no. 4, pp. 2139–2147, Apr. 2014.

[3] A. Samal and S. Jain, "The regulatory network of E. Coli metabolism as a Boolean dynamical system exhibits both homeostasis and flexibility of response," *BMC Syst. Biol.*, vol. 2, no. 1, p. 21, 2008.

[4] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," in *Proc. 37th Annu. Allerton Conf. Commun., Control, Comput.*, Sep. 1999, pp. 368–377.

[5] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. New York, NY, USA: Wiley, 2006.

[6] A. D. Wyner and J. Ziv, "A theorem on the entropy of certain binary sequences and applications (part 1)," *IEEE Trans. Inf. Theory*, vol. 19, no. 6, pp. 769–772, Nov. 1973.

[7] E. Erkip, "The efficiency of information in investment," Ph.D. dissertation, Dept. Electr. Eng., Stanford Univ. Press, Stanford, CA, USA, 1996.

[8] R. O'Donnell, "Some topics in analysis of Boolean functions," in *Proc. 40th Annu. ACM Symp. Theory Comput.*, 2008, pp. 569–578.

[9] E. Friedgut and G. Kalai, "Every monotone graph property has a sharp threshold," *Proc. Amer. Math. Soc.*, vol. 124, no. 10, pp. 2993–3002, 1996.

[10] L. Harper, "Optimal numberings and isoperimetric problems on graphs,," *J. Combinat. Theory*, no. 1, no. 3, pp. 385–393, 1966.

[11] B. Bollobás and I. Leader, "Compressions and isoperimetric inequalities," *J. Combinat. Theory, Ser. A*, vol. 56, no. 1, pp. 47–62, 1991.

[12] P. C. Allaart and K. Kawamura, "The Takagi function: A survey," *Real Anal. Exchange*, vol. 37, no. 1, pp. 1–54, 2011.

[13] C. J. Guu, "The McFunction," *Discrete Math.*, vol. 213, nos. 1–3, pp. 163–167, Feb. 2000.

[14] V. Anantharam, A. A. Gohari, S. Kamath, and C. Nair, "On hypercontractivity and the mutual information between Boolean functions," in *Proc. 51st Annu. Allerton Conf. Commun., Control, Comput.*, Oct. 2013, pp. 13–19.

[15] C. Nair, private communication, Jul. 2013.

[16] F. d. P. Calmon, M. Varia, and M. Médard, "An exploration of the role of principal inertia components in information theory," arXiv:1405.1472 [cs.IT], to be published.

[17] V. Chandar and A. Tchamkerten, "Most informative quantization functions (presentation)," in *Proc. ITA Workshop*, San Diego, CA, USA, Feb. 2014.

[18] D. Ellis, "Almost isoperimetric subsets of the discrete cube," *Combinat., Probab. Comput.*, vol. 20, no. 3, pp. 363–380, 2011.

[19] T. Takagi, "A simple example of a continuous function without derivative," *Proc. Phys. Math. Jpn.*, vol. 1, pp. 176–177, 1903.

[20] J. R. Trollope, "An explicit expression for binary digital sums," *Math. Mag.*, vol. 41, no. 1, pp. 21–25, Jan. 1968.

[21] R. Ahlswede and N. Cai, "General edge-isoperimetric inequalities, part I: Information-theoretical methods," *Eur. J. Combinat.*, vol. 18, no. 4, pp. 355–372, 1997.

[22] R. Ahlswede and N. Cai, "General edge-isoperimetric inequalities, part II: A local–global principle for lexicographical solutions," *Eur. J. Combinat.*, vol. 18, no. 5, pp. 479–489, 1997.

[23] Y. Crama and P. L. Hammer, *Boolean Functions: Theory, Algorithms, and Applications* (Encyclopedia of Mathematics and Its Applications). Cambridge, U.K.: Cambridge Univ. Press, 2011.

Thomas A. Courtade is an Assistant Professor in the Department of Electrical Engineering and Computer Sciences at the University of California, Berkeley. Prior to joining UC Berkeley in 2014, he was a postdoctoral fellow supported by the NSF Center for Science of Information. He received his Ph.D. and M.S. degrees from UCLA in 2012 and 2008, respectively, and he graduated summa cum laude with a B.Sc. in Electrical Engineering from Michigan Technological University in 2007.

His honors include a Distinguished Ph.D. Dissertation award and an Excellence in Teaching award from the UCLA Department of Electrical Engineering, and a Jack Keil Wolf Student Paper Award for the 2012 International Symposium on Information Theory.

Gowtham R. Kumar is a Ph.D. student in the Department of Electrical Engineering at Stanford University, working under the supervision of Professor Abbas El Gamal. He received his M.S. degree from Stanford University in 2010 and his B.Tech. degree from IIT Madras in 2008. His research interests include Information Theory, Statistics, and Gambling.

He is the recipient of a departmental fellowship from Stanford University, and an Indian National Mathematics Olympiad (INMO) awardee.