Comments and Corrections

Comments on "Canalizing Boolean Functions Maximize Mutual Information"

Thomas A. Courtade, *Member, IEEE*

***Abstract*— In their recent paper "Canalizing Boolean Functions Maximize Mutual Information," Klotz *et al.* argued that canalizing Boolean functions maximize certain mutual informations by an argument involving Fourier analysis on the hypercube. This note supplies short new proofs of their results based on a coupling argument and also clarifies a point on the necessity of considering randomized functions.**

***Index Terms*— Boolean functions, mutual information.**

A (possibly randomized) Boolean function $f : \{0,1\}^n \to \{0,1\}$ is said to be *canalizing* in x_i if there exists $v \in \{0,1\}$ such that $f(x^n)$ is constant for all x^n having $x_i = v$. These functions are of interest due in part to their relevance to genetic network models [1]. In a recent paper, Klotz *et al.* showed that canalizing Boolean functions maximize certain mutual informations by way of an argument involving Fourier analysis on the hypercube [2]. Specifically, Klotz *et al.* established necessary and sufficient conditions for a Boolean function g to maximize mutual information subject to a bias constraint in the following problem:

$$I(g(X^n); X_i) = \max_{f:\mathbb{E}f(X^n)=\mathbb{E}g(X^n)} I(f(X^n); X_i), \qquad (1)$$

where the maximum is taken over all (possibly randomized) Boolean functions f with bias $\mathbb{E}[f(X^n)]$ equal to that of g.

In this note, we give short new proofs of their results using a simple coupling lemma as our primary tool[1] and also clarify a point on the necessity of considering randomized functions in (1). For reference, we now recall the results of [2].

Theorem 1 [2, Th. 1,2]: Fix $i \in \{1,\dots,n\}$. Let X^n be arbitrarily distributed on $\{0,1\}^n$, and suppose $0 < \mathbb{E}[X_i]$, $\mathbb{E}[g(X^n)] < 1/2$ for a Boolean function g.

i) If $\mathbb{E}[X_i] \geq \mathbb{E}[g(X^n)]$, then (1) holds iff g is canalizing in x_i with $g(x^n) = 0$ whenever $x_i = 0$.

ii) If $\mathbb{E}[X_i] \leq \mathbb{E}[g(X^n)]$, then (1) holds iff g is canalizing in x_i with $g(x^n) = 1$ whenever $x_i = 1$.

In Theorem 1, we exclude $\min\{\mathbb{E}[X_i], \mathbb{E}[g(X^n)]\} = 0$ to avoid degeneracy: if $\mathbb{E}[g(X^n)] = 0$, then g is trivially canalizing in all variables; if $\mathbb{E}[X_i] = 0$ then $X_i = 0$ with probability one, implying that $I(f(X^n); X_i) = 0$ for any function f, not only those that are canalizing.

The case where $\max\{\mathbb{E}[X_i], \mathbb{E}[g(X^n)]\} = 1/2$ is ambiguous in [2, Th. 2] due to $\mathsf{sign}(0)$ being undefined, and is also excluded from Theorem 1 above. Nevertheless, it is possible to deduce the following result:

Theorem 2 [2, Th. 1,2]: Fix $i \in \{1,\dots,n\}$. Let X^n be arbitrarily distributed on $\{0,1\}^n$, and suppose $\min\{\mathbb{E}[X_i], \mathbb{E}[g(X^n)]\} > 0$ and $\max\{\mathbb{E}[X_i], \mathbb{E}[g(X^n)]\} = 1/2$ for a Boolean function g.

i) If $\mathbb{E}[X_i] \geq \mathbb{E}[g(X^n)]$, then (1) holds iff g is canalizing in x_i with $g(x^n) = 0$ whenever $x_i = v$ for some $v \in \{0,1\}$.

ii) If $\mathbb{E}[X_i] \leq \mathbb{E}[g(X^n)]$, then (1) holds iff g is canalizing in x_i with $g(x^n)$ being constant for all x^n having $x_i = 1$.

Remark 1: The constraint $\max\{\mathbb{E}[X_i], \mathbb{E}[g(X^n)]\} \leq 1/2$ is not present in [2], but does not sacrifice generality in the stated versions of Theorems 1 and 2. Indeed, mutual information is invariant to relabeling, so complementing X_i and/or $g(X^n)$ does not change the value of $I(g(X^n); X_i)$.

When interpreting Theorems 1 and 2 (and the results that ensue), it is necessary to consider randomized functions. This means that functions need not depend on their inputs in a deterministic manner. In other words, a randomized function $f(x^n)$ is a binary random variable with distribution that depends on x^n:

$$f(x^n) = \begin{cases} 0 & \text{with probability } q(x^n) \\ 1 & \text{with probability } 1 - q(x^n). \end{cases} \qquad (2)$$

Note that the definition of 'canalizing' applies equally to deterministic or randomized functions. In particular, a randomized function f that is canalizing in x_i satisfies $P\{f(X^n) = u | X_i = v\} = 1$ for some $u, v \in \{0,1\}$. That is, $x_i = v$ implies $f(x^n) = u$.

Although not addressed in their paper [2], Klotz *et al.* must permit randomization in their formulation. In fact, the results in [2] do not hold if only deterministic functions are considered, as illustrated by the following counterexample:

Example 1: Let $P\{X^2 = x^2\} = p_1(x_1)p_2(x_2)$, where $p_1(1) = 1 - p_1(0) = 1/5$ and $p_2(1) = 1 - p_2(0) = 2/5$. Consider the function $g : \{0,1\}^2 \to \{0,1\}$ defined by:

x_1x_2	00	01	10	11
$g(x^2)$	0	1	1	0

By inspection, $g(x^2)$ is not canalizing in either variable. Furthermore, $\mathbb{E}[g(X^2)] = 11/25$, and is the only deterministic Boolean function on $\{0,1\}^2$ with this expected value under the

Manuscript received May 17, 2014; revised September 4, 2014; accepted November 4, 2014. Date of publication November 26, 2014; date of current version January 16, 2015.

The author is with the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, Berkeley, CA 94720 USA (e-mail: courtade@berkeley.edu).

Communicated by O. Milenkovic, Associate Editor for Coding Theory.

Digital Object Identifier 10.1109/TIT.2014.2375183

[1] In fact, we generalize the results in [2] since Klotz *et al.* restrict their attention to product distributions on $\{0,1\}^n$.

specified distribution. Hence, even though

$$I(g(X^n); X_1) = \max_{f:\mathbb{E}f(X^n)=\mathbb{E}g(X^n)} I(f(X^n); X_1) \qquad (3)$$

when the maximization is restricted to deterministic functions, we can not assert that g is canalizing.

However, when X^n assumes the uniform distribution, restriction to deterministic functions is possible since if $k = |\{x^n : f(x^n) = 1\}| \leq \frac{1}{2}2^n$ for a deterministic function f, there always exists a deterministic canalizing function g with $k = |\{x^n : g(x^n) = 1\}|$, implying $\mathbb{E}[g(X^n)] = \mathbb{E}[f(X^n)] = k2^{-n}$.

We now show that Theorems 1 and 2 are consequences of the following coupling lemma, which is proved in the Appendix.

Lemma 1: Suppose U, V are binary random variables with $0 < \mathbb{E}[U] \leq \mathbb{E}[V] \leq 1/2$. There exists a random variable $\widetilde{V}$ with $\mathbb{E}[\widetilde{V}] = \mathbb{E}[V]$ such that $P\{\widetilde{V} = 1|U = 1\} = 1$ and $I(U; \widetilde{V}) \geq I(U; V)$, with equality iff $\widetilde{V} = V$ or $\widetilde{V} = 1 - V$.

Remark 2: Note that the second sufficient condition for equality in Lemma 1 (i.e, that $\widetilde{V} = 1 - V$) can only hold when $\mathbb{E}[V] = 1/2$. Indeed, $\widetilde{V} = 1 - V$ implies $\mathbb{E}[\widetilde{V}] = 1 - \mathbb{E}[V]$, and therefore $\mathbb{E}[V] = 1/2$, since $\mathbb{E}[\widetilde{V}] = \mathbb{E}[V]$ by definition.

In words, Lemma 1 couples U with a new random variable $\widetilde{V}$ equal in distribution to V that *i*) improves mutual information in the sense that $I(U; \widetilde{V}) \geq I(U; V)$; and *ii*) is 'canalizing in U' in the sense that $\{U = 1\} \Rightarrow \{\widetilde{V} = 1\}$.

With Lemma 1 in hand, Theorems 1 and 2 are readily proved by identifying appropriate assignments of U, V.

Proof of Theorems 1 and 2: We consider two cases.

▹ *Case 1:* $\mathbb{E}[X_i] \geq \mathbb{E}[g(X^n)]$. Define $U = g(X^n)$ and $V = X_i$. Applying Lemma 1, there exists $\widetilde{V}$ equal to V in distribution, satisfying $P\{\widetilde{V} = 1|U = 1\} = 1$ and $I(U; \widetilde{V}) \geq I(U; V)$, with equality iff V is in one-to-one correspondence with $\widetilde{V}$. If $I(U; \widetilde{V}) > I(U; V)$, then optimality of g in (1) is contradicted, since taking $q(x^n) = P\{U = 0|\widetilde{V} = x_i\}$ in (2) defines a randomized function f which satisfies $I(f(X^n); X_i) > I(g(X^n); X_i)$ because $(f(X^n), X_i)$ equals $(U, \widetilde{V})$ in distribution.

Therefore, assume $I(U; \widetilde{V}) = I(U; V)$. Supposing first that $\mathbb{E}[X_i] < 1/2$, Lemma 1 and the remark that follows assert that $P\{X_i = 1|g(X^n) = 1\} = 1$. By definition, this implies that $P\{g(X^n) = 0|X_i = 0\} = 1$, so g is canalizing in x_i as stated. On the other hand, if $\mathbb{E}[X_i] = 1/2$, then we must have $P\{X_i = v|g(X^n) = 1\} = 1$ for some $v \in \{0, 1\}$, which implies that $P\{g(X^n) = 0|X_i = 1 - v\} = 1$.

▹ *Case 2:* $\mathbb{E}[X_i] \leq \mathbb{E}[g(X^n)]$. Reversing roles in the previous case, we define $V = g(X^n)$ and $U = X_i$ and apply Lemma 1. Arguing as above, if g satisfies (1), then we must have $P\{g(X^n) = 1|X_i = 1\} = 1$ when $\mathbb{E}[g(X^n)] < 1/2$, showing g is canalizing in x_i. On the other hand, if $\mathbb{E}[g(X^n)] = 1/2$, then we must have $P\{g(X^n) = v|X_i = 1\} = 1$ for some $v \in \{0, 1\}$. □

By the data processing inequality, $I(g(X^n); X_i) \leq H(X_i)$ with equality iff $g(x^n)$ is a dictatorship in x_i (i.e., a one-to-one function of x_i) assuming that X_i is nondegenerate. Hence, we also have:

Theorem 3 [2, Th. 3]: Let X^n be arbitrarily distributed on $\{0, 1\}^n$ with X_i not a.s. constant. A Boolean function g is a dictatorship in x_i iff $I(g(X^n); X_i) = \max_f I(f(X^n); X_i)$.

A Boolean function f is said to be *jointly canalizing* in $T \subset [n]$ if there exists $\{v_i\}_{i\in T} \in \{0, 1\}^{|T|}$ such that $f(x^n)$ is constant for all x^n having $x_i = v_i$ for $i \in T$. Define $X_T = \{X_i : i \in T\}$. A modification of the proof of Theorem 1 gives:

Theorem 4 [2, Th. 4]: Let X^n be arbitrarily distributed on $\{0, 1\}^n$. If

$$I(g(X^n); X_T) = \max_{f:\mathbb{E}f(X^n)=\mathbb{E}g(X^n)} I(f(X^n); X_T) > 0, \quad (4)$$

then g is jointly canalizing in T.

Proof: Let $k = |T|$. By the assumption that $I(g(X^n); X_T) > 0$, neither $g(X^n)$ nor X_T are constant. Therefore, there must exist two sequences $y^k, z^k \in \{0, 1\}^k$ such that $0 < P\{X_T = y^k\} \leq P\{X_T = z^k\} < 1$. Define $A = \{y^k, z^k\}$, and observe that $0 < P\{X_T = y^k|X_T \in A\} \leq 1/2$. Further, by complementing $g(X^n)$ if necessary, we may assume without loss of generality that $P\{g(X^n) = 1|X_T \in A\} \leq 1/2$. If $P\{g(X^n) = 1|X_T \in A\} = 0$, then g is canalizing in T as desired. Therefore, we consider $0 < P\{g(X^n) = 1|X_T \in A\} \leq 1/2$.

Conditioning on the event $\{X_T \in A\}$, we repeat the argument in the proof of Theorem 1 letting $V = \mathbf{1}\{X_T = y^k\}$ and $U = g(X^n)$, or letting $U = \mathbf{1}\{X_T = y^k\}$ and $V = g(X^n)$, depending on whether $P\{X_T = y^k|X_T \in A\}$ or $P\{g(X^n) = 1|X_T \in A\}$ is larger. By doing so, we can conclude that $g(x^n)$ is constant when either $x_T = y^k$ or $x_T = z^k$. The details are the same as in the proof of Theorem 1 and are left to the reader. □

Concluding Remarks

In this note, we recovered the results of Klotz *et al.* in [2] by replacing the Fourier-analytic machinery they employed with an alternative argument based on coupling and data processing. In proceeding along this route, we avert the need to restrict our attention to product distributions on X^n as was required in the Fourier-based proof of Klotz *et al.* On reflection, it should not be surprising that Fourier analysis isn't essential to the proofs since the mutual information $I(g(X^n); X_i)$ only depends on the joint distribution of the binary pair $(g(X^n), X_i)$, a fact emphasized in the present paper.

Appendix

Proof of Lemma 1: To simplify notation, define $\bar{\mu} = 1 - \mu$ for $\mu \in [0, 1]$. Also, let $q = \mathbb{E}V$, $p = \mathbb{E}U$ and $\epsilon = P\{V = 0|U = 1\}$. Observe that p, q, ϵ completely specify the joint distribution of U, V. Next, define $t = (q - p)/\bar{p}$, which is in $[0, 1]$ since $p \leq q \leq 1$. Finally, define $\delta = \epsilon q/\bar{q}$, which is also in $[0, 1]$ since $q \leq 1/2 \leq \bar{q}$. Using these definitions, we observe that the cascade shown in Figure 1 induces the correct joint distribution on U, V. Indeed, for $U \sim$ Bernoulli(p) input to the cascade, we see that the V has the desired marginal distribution:

$$\begin{aligned} P\{V = 1\} &= (1 - \epsilon)P\{\widetilde{V} = 1\} + \delta P\{\widetilde{V} = 0\} \\ &= (1 - \epsilon)(p + t(1 - p)) + \delta(1 - t)(1 - p) = q. \end{aligned}$$

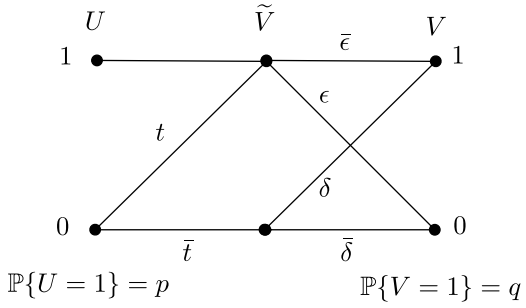

Fig. 1. A cascade relating U and V when $\mathbb{E}U \leq \mathbb{E}V \leq 1/2$.

Moreover, the probability of going from $U = 1$ to $V = 0$ in the cascade is ϵ, which equals $P\{V = 0|U = 1\}$ by definition.

Taking $\widetilde{V}$ to be the intermediate variable in the cascade implies that $U \to \widetilde{V} \to V$ form a Markov chain, in that order. Thus, the data processing inequality implies that $I(U; \widetilde{V}) \geq I(U; V)$, with equality iff V is a one-to-one function of $\widetilde{V}$. Clearly $P\{\widetilde{V} = 1|U = 1\} = 1$, so the proof is complete by noting that $\widetilde{V}$ is equal in distribution to V as desired since

$$P\{\widetilde{V} = 1\} = P\{U = 1\} + tP\{U = 0\} = p + \tfrac{q-p}{\bar{p}}\bar{p} = q.$$

□

ACKNOWLEDGMENT

The author would like to thank the anonymous reviewers whose comments substantially improved this note.

REFERENCES

[1] S. Kauffman, C. Peterson, B. Samuelsson, and C. Troein, "Genetic networks with canalyzing Boolean rules are always stable," *Proc. Nat. Acad. Sci. United States Amer.*, vol. 101, no. 49, pp. 17102–17107, 2004.

[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.

Thomas A. Courtade (S'06–M'13) 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.