Bounds on the Mutual Informations of the Binary Sums of Bernoulli Random Variables

Payam Pakzad
Qualcomm Inc.
Santa Clara, CA 95051
payam@qualcomm.com

Venkat Anantharam
EECS Department, U.C.Berkeley
Berkeley, CA 94720
ananth@eecs.berkeley.edu

Amin Shokrollahi
Algorithmics Laboratory, EPFL
Lausanne, Switzerland
amin.shokrollahi@epfl.ch

***Abstract*—We present some simple information inequalities on binary sums of Bernoulli random variables that appear to be new. Consequences for information across binary input memoryless symmetric channels are also presented.**

I. Setup and Preliminary Results

Consider a collection of independent Bernoulli random variables $\{B_1, \ldots, B_k\}$. Let '$\oplus$' denote the binary sum operation. If S is a binary sum of the Bernoulli variables B_i, $1 \leq i \leq k$, i.e. $S := \bigoplus_{i=1}^{k} g_i B_i$ where $g_i \in \{0, 1\}$, then S will be called a *BSBV* variable and we will refer to the B_i's as the *base* variables in the definition of S.

For each integer m, define the following function on $[0, 1]^m$:

$$S(p_1, \ldots, p_m) := \prod_{i=1}^{m} p_i + \prod_{i=1}^{m} (1 - p_i) \ . \tag{1}$$

Then $S(p_1, \ldots, p_m)$ can be interpreted as the probability that all BSBVs comprised of pairwise binary sums on m base variables with parameters $p_1, \ldots, p_m$ respectively are simultaneously 0.

Our first result is the following, which, to the best of our knowledge, is new:

Lemma 1: Let E, X and Y be independent Bernoulli random variables with parameters $\lambda := P(E = 1), p := P(X = 1)$, $q := P(Y = 1)$, where $p, q \leq \frac{1}{2}$. Then the mutual information $I(E \oplus X \ ; \ E \oplus Y)$ is a non-decreasing function of λ for $0 \leq \lambda \leq \frac{1}{2}$.

Proof: We demonstrate that the derivative of the mutual information term in question w.r.t. λ is non-negative for all $\lambda \in [0, \frac{1}{2}]$.

Let $p(i)$ denote $P(X = i)$ and $q(j)$ denote $P(Y = j)$, where $i, j \in \{0, 1\}$. Also, for real valued $0 \leq u, v \leq 1$ define

$$F(u, v) := \log \frac{S(\lambda, u, v)}{S(\lambda, u)\, S(\lambda, v)},$$

where $S(.)$ is as defined in (1). Then we have

$$\begin{aligned} \mathcal{D} &:= \frac{\partial}{\partial \lambda} I(E \oplus X \ ; \ E \oplus Y) \\ &= \frac{\partial}{\partial \lambda} \sum_{i,j \in \{0,1\}} S(\lambda, p(i), q(j)) \log \frac{S(\lambda, p(i), q(j))}{S(\lambda, p(i))\, S(\lambda, q(j))} \\ &= \sum_{i,j} \big(p(i)q(j) - (1 - p(i))(1 - q(j))\big) \cdot F\big(p(i), q(j)\big) \end{aligned} \tag{2}$$

where in the last equality, all the terms obtained by taking derivative of the terms under the logarithm sum to zero. Next, combining the similar terms we get the following expression for the derivative:

$$\begin{aligned} \mathcal{D} &= \big((1 - p)(1 - q) - p\, q\big)\big[F(1 - p, 1 - q) - F(p, q)\big] \\ &\quad + \big(p(1 - q) - (1 - p)q\big)\big[F(p, 1 - q) - F(1 - p, q)\big] \\ &= (1 - p - q)\big[F(1 - p, 1 - q) - F(p, q)\big] \\ &\quad + (p - q)\big[F(p, 1 - q) - F(1 - p, q)\big] \end{aligned} \tag{3}$$

We will now examine the signs of each of the four terms involved in (3). We will assume for convenience that $p \leq q \leq \frac{1}{2}$, although the other combinations can be treated in a similar fashion. Then $(p - q) \leq 0$ and $(1 - p - q) \geq 0$.

Next, expanding out the terms for the expression $[F(1 - p, 1 - q) - F(p, q)]$ in (3), we get

$$\begin{aligned} &F(1 - p, 1 - q) - F(p, q) \\ &\quad = \log \frac{S(\lambda, 1 - p, 1 - q)\, S(\lambda, p)\, S(\lambda, q)}{S(\lambda, p, q)\, S(\lambda, 1 - p)\, S(\lambda, 1 - q)} \\ &\quad = \log \frac{\lambda^3 a + \lambda^2(1 - \lambda)b + \lambda(1 - \lambda)^2 c + (1 - \lambda)^3 a}{\lambda^3 a + \lambda^2(1 - \lambda)c + \lambda(1 - \lambda)^2 b + (1 - \lambda)^3 a} \end{aligned} \tag{4}$$

for appropriate a, b, c, d where $c - b = (1 - p - q)(1 - 2p)(1 - 2q) \geq 0$. It then follows that, for $\lambda \leq \frac{1}{2} \leq (1 - \lambda)$ the numerator is greater than or equal to the denominator, and hence the expression (4) is non-negative.

Using transformation $p \to (1 - p)$ in (4) we similarly obtain

$$\begin{aligned} &F(p, 1 - q) - F(1 - p, q) \\ &\quad = \log \frac{\lambda^3 a' + \lambda^2(1 - \lambda)b' + \lambda(1 - \lambda)^2 c' + (1 - \lambda)^3 a'}{\lambda^3 a' + \lambda^2(1 - \lambda)c' + \lambda(1 - \lambda)^2 b' + (1 - \lambda)^3 a'} \end{aligned}$$

where $c' - b' = (q - p)(1 - 2p)(1 - 2q) \geq 0$, proving that $F(p, 1 - q) - F(1 - p, q) \geq 0$.

Finally, since $(1-p-q) \geq (q-p)$, in order to show that $\mathcal{D} \geq 0$ in (3), it suffices to show that

$$\big[F(1-p,1-q)-F(p,q)\big]-\big[F(p,1-q)-F(1-p,q)\big] \geq 0 \quad (5)$$

Once again we expand the terms in (5), cancelling out the identical terms:

$$\begin{aligned} &F(1-p,1-q) - F(p,q) - F(p,1-q) + F(1-p,q) \\ &\quad = \log \frac{S(\lambda,1-p,1-q)\, S(\lambda,1-p,q)\, S(\lambda,p)^2}{S(\lambda,p,q)\, S(\lambda,p,1-q)\, S(\lambda,1-p)^2} \\ &\quad = \log \frac{(\phi+\gamma\,\tau)(\psi+2\tau)}{(\psi+\gamma\,\tau)(\phi+2\tau)} \end{aligned} \quad (6)$$

where

$$\begin{aligned} \phi &= \lambda^2(1-p)^2 + (1-\lambda)^2 p^2, \\ \psi &= \lambda^2 p^2 + (1-\lambda)^2(1-p)^2, \\ \tau &= \lambda(1-\lambda)p(1-p), \qquad \text{and} \\ \gamma &= \frac{q^2+(1-q)^2}{q(1-q)} \end{aligned}$$

It can then be seen that $\psi - \phi = (1-2\lambda)(1-2p) \geq 0$, and $\tau \geq 0$, and $\gamma \geq 2$. It follows that

$$(\phi+\gamma\,\tau)(\psi+2\tau) - (\psi+\gamma\,\tau)(\phi+2\tau) = (\gamma-2)(\psi-\phi) \geq 0$$

The fraction under the logarithm in (6) is thus bounded below by unity, and hence the expression in (5) is non-negative, as required. This completes the proof. ∎

Regarding the statement of Lemma 1, note that the restriction on $p, q \leq \frac{1}{2}$ is in fact unnecessary, as clearly the mutual information is unchanged if X is replaced by $1-X$ (or Y by $1-Y$). It can also be easily verified that this mutual information is a symmetric function of λ around the point $\lambda = \frac{1}{2}$.

A slightly more general version of the result of Lemma 1 can be immediately derived:

Lemma 2: Let A and B be BSBV variables over the base $\{E, X_1, \ldots, X_m\}$, and let $\lambda := P(E=1)$ as before. Then the mutual information $I(A\,;\,B)$ is a monotonic function of λ for $0 \leq \lambda \leq \frac{1}{2}$.

Proof: Depending on how the variable E is involved in the linear combinations that determine A and B, we will have three scenarios:

If E does not appear in neither of A and B, then $I(A;B)$ is constant with respect to λ, and the statement is vacuous.

If E is involved in both, then we must have $A = E \oplus F \oplus X$, and $B = E \oplus F \oplus Y$, where F contains the common base variables appearing in both A and B, and X and Y contains the distinct variables. Then, with $E' := E \oplus F$, it follows from the previous lemma that $I(A;B) = I(E' \oplus X\,;\,E' \oplus Y)$ is a non-decreasing function of $P(E'=1) = S(\lambda, P(F=0))$. Now $S(\lambda, P(F=0))$ is itself a monotonic function of $\lambda \in [0, \frac{1}{2}]$, which proves the statement.

If E appears in exactly one of A and B, say $A = E \oplus F \oplus X$, and $B = F \oplus Y$, then $I(A;B) = H(B) - H(B|A)$. The first term does not depend on λ, and the second term $H(F \oplus Y | E \oplus F \oplus X)$ is a monotonically decreasing function of λ for $\lambda \in [0, \frac{1}{2}]$. This completes the proof. ∎

Consider the problem of transmission of a codeword of a binary linear code over a memoryless binary symmetric channel (BSC); then each noisy observation corresponds to a BSBV, where the base variables consist of the original source bits and the BSC noise variables. With this viewpoint the preceding results can be generalized to *binary input memoryless symmetric channels*, defined as follows:

Definition 1: A Binary Input Memoryless Symmetric Channel (BIMSC) with (countable) output alphabet $\mathcal{A}$ is a channel $\mathcal{C}$ with a binary input b and (random) output $\mathcal{C}(b) \in \mathcal{A}$ such that for every $x \in \mathcal{A}$, there exists a $y \in \mathcal{A}$, denoted by $y = \hat{x}$, such that $\hat{y} = x$, and $P(\mathcal{C}(0) = x) = P(\mathcal{C}(1) = \hat{x})$.

Note from the above definition that the vector channel created from finitely many independent BIMSCs is itself a BIMSC. To see this, suppose $\mathcal{C}_1$ and $\mathcal{C}_2$ are two independent BIMSCs with output alphabets $\mathcal{A}_1$ and $\mathcal{A}_2$ respectively. The vector channel $\mathcal{C}_{1,2} := \big(\mathcal{C}_1, \mathcal{C}_2\big)$, is defined to have output $\mathcal{C}_{1,2}(X) := \big(\mathcal{C}_1(X), \mathcal{C}_2(X)\big) \in \mathcal{A}_{1,2} := \mathcal{A}_1 \times \mathcal{A}_2$. Then for every $z = (x,y) \in \mathcal{A}_{1,2}$, and with $\hat{z} := (\hat{x}, \hat{y})$ we have

$$\begin{aligned} P\big(\mathcal{C}_{1,2}(0) = z\big) &= P\big(\mathcal{C}_1(0) = x, \mathcal{C}_2(0) = y\big) \\ &= P\big(\mathcal{C}_1(0) = x\big) \cdot P\big(\mathcal{C}_2(0) = y\big) \\ &= P\big(\mathcal{C}_1(1) = \hat{x}\big) \cdot P\big(\mathcal{C}_2(1) = \hat{y}\big) \\ &= P\big(\mathcal{C}_{1,2}(1) = \hat{z}\big) \end{aligned}$$

where we have used the independence and the BIMSC properties of the $\mathcal{C}_1$ and $\mathcal{C}_2$. This shows that $\mathcal{C}_{1,2}$ is a BIMSC. The argument extends to the vector channel created from finitely many BIMSCs by induction.

We now generalize the result of Lemma 2:

Proposition 3: Let E be a Bernoulli(λ) random variable, and let $\mathcal{C}_1$ and $\mathcal{C}_2$ denote two independent BIMSCs with output alphabet $\mathcal{A}_1$ and $\mathcal{A}_2$, respectively. Then the mutual information $I(\mathcal{C}_1(E)\,;\,\mathcal{C}_2(E))$ is a non-decreasing function of λ for $0 \leq \lambda \leq \frac{1}{2}$.

Proof: For convenience, we define the following functions: $p_i(x) := P(\mathcal{C}_i(0) = x)$, $\pi_i(x) := \pi_i(\hat{x}) := p_i(x) + p_i(\hat{x})$, and $\bar{p}_i(x) := \frac{p_i(x)}{\pi_i(x)}$. It follows that $\pi_i(x) = \pi_i(\hat{x})$, and

$\bar{p}_i(\hat{x}) = 1 - \bar{p}_i(x)$. Then we have

$$
\begin{aligned}
&I(\mathcal{C}_1(E)\,;\,\mathcal{C}_2(E)) \\
&\quad= \sum_{x\in\mathcal{A}_1, y\in\mathcal{A}_2} P(\mathcal{C}_1 = x, \mathcal{C}_2 = y) \log \frac{P(\mathcal{C}_1 = x, \mathcal{C}_2 = y)}{P(\mathcal{C}_1 = x)P(\mathcal{C}_2 = y)} \\
&\quad= \sum_{x,y} \big(\lambda\, p_1(x)p_2(y) + (1-\lambda)p_1(\hat{x})p_2(\hat{y})\big)\cdot \\
&\qquad \log \frac{\lambda\, p_1(x)p_2(y) + (1-\lambda)p_1(\hat{x})p_2(\hat{y})}{\big(\lambda\, p_1(x) + (1-\lambda)p_1(\hat{x})\big)\big(\lambda\, p_2(y) + (1-\lambda)p_2(\hat{y})\big)} \\
&\quad= \sum_{x,y} \pi_1(x)\pi_2(y)\cdot \\
&\qquad S\big(\lambda, \bar{p}_1(x), \bar{p}_2(y)\big) \cdot \log \frac{S\big(\lambda, \bar{p}_1(x), \bar{p}_2(y)\big)}{S\big(\lambda, \bar{p}_1(x)\big)S\big(\lambda, \bar{p}_2(y)\big)} \\
&\quad= \frac{1}{4}\sum_{x,y} \pi_1(x)\pi_2(y)\, I(E\oplus N_x\,;\, E\oplus N_y) \qquad (7)
\end{aligned}
$$

where N_x and N_y are (dummy) independent Bernoulli random variables with parameters $\bar{p}_1(x)$ and $\bar{p}_2(y)$: in the second equality we have used the fact that conditioned on the value of E, $\mathcal{C}_1(E)$ and $\mathcal{C}_2(E)$ are independent; the third equality follows from definitions; to see the last equality, note that for each pair (x, y) with $\hat{x} \neq x$ and $\hat{y} \neq y$, the summand in the final expression represents four distinct terms, corresponding to $(x, y), (\hat{x}, y), (x, \hat{y})$ and $(\hat{x}, \hat{y})$. If on the other hand $x = \hat{x}$ (or $y = \hat{y}$), then $p_1(x) = p_1(\hat{x})$ and hence $\bar{p}_1(x) = \frac{1}{2}$, therefore $I(E\oplus N_x\,;\, E\oplus N_y) = 0$, which does not contribute to the sum.

The claim now follows, since by Lemma 1 each term under the sum is a non-decreasing function of λ in $[0, \frac{1}{2}]$. ∎

Equation (7) is an example of a useful technique in dealing with arbitrary BIMSCs, when it is possible to reduce the problem as an appropriately weighted superposition of BSCs with different parameters.

II. An Upper bound on the Mutual Information of Sums of Independent Random Variables

We start with the simpler case of independent Bernoulli variables:

Proposition 4: Let X be a Bernoulli($\frac{1}{2}$) random variable, and $N_0, N_1, \ldots, N_m$ be independent Bernoulli variables. Then the following inequality holds:

$$I(X\oplus N_0\,;\, X\oplus N_1, \ldots, X\oplus N_m) \leq \sum_{i=1}^{m} I(X\oplus N_0\,;\, X\oplus N_i) \qquad (8)$$

Proof: The statement is clearly correct for $m = 1$. To prove for the case $m > 1$ we proceed by induction. We first expand the mutual information as follows:

$$
\begin{aligned}
&I(X\oplus N_0\,;\, X\oplus N_1, \ldots, X\oplus N_m) \\
&\quad= H(X\oplus N_0) - H(X\oplus N_0 \mid X\oplus N_1, \ldots, X\oplus N_m) \\
&\quad= 1 - H(N_0\oplus N_1 \mid X\oplus N_1, N_1\oplus N_2, \ldots, N_1\oplus N_m) \\
&\quad= 1 - H(N_0\oplus N_1 \mid N_1\oplus N_2, \ldots, N_1\oplus N_m) \qquad (9)
\end{aligned}
$$

where the second equality follows from the fact that X is an independent Bernoulli($\frac{1}{2}$) variable, and where we have added the term $X\oplus N_1$ to all other terms; the third equality follows from the fact that $X\oplus N_1$ is independent of $(N_0, N_1, \ldots, N_m)$.

Similarly we can show that for each $i \in \{1, \ldots, m\}$

$$I(X\oplus N_0\,;\, X\oplus N_i) = 1 - H(N_0\oplus N_i) \qquad (10)$$

Applying (9) and (10) to (8), we get

$$
\begin{aligned}
D &:= I(X\oplus N_0\,;\, X\oplus N_1, \ldots, X\oplus N_m) - \\
&\qquad \sum_{i=1}^{m} I(X\oplus N_0\,;\, X\oplus N_i) \\
&= 1 - H(N_0\oplus N_1 \mid N_1\oplus N_2, \ldots, N_1\oplus N_m) - \\
&\qquad \sum_{i=1}^{m} \big(1 - H(N_0\oplus N_i)\big) \\
&= I(N_0\oplus N_1\,;\, N_1\oplus N_2, \ldots, N_1\oplus N_m) - \\
&\qquad \sum_{i=2}^{m} \big(1 - H(N_0\oplus N_i)\big) \qquad (11)
\end{aligned}
$$

where in the last equality we have combined the first term of the sum with the conditional entropy to obtain the mutual information term.

Now for fixed distribution of $(N_0, N_2, \ldots, N_m)$, the expression in (11) is only a function of $p_1 = P(N_1 = 1)$ through the mutual information term. Next note that the combination of the variables $(N_1\oplus N_2, \ldots, N_1\oplus N_m)$ is equivalent to a noisy observation of N_1 over a BIMSC with output alphabet $\mathrm{GF}(2)^{m-1}$. Using Proposition 3 then (11) is a non-decreasing function of p_1 in $[0, \frac{1}{2}]$. It thus suffices to show that D in (11) is less than or equal to zero for $p_1 = \frac{1}{2}$. But with $p_1 = \frac{1}{2}$, this condition is simply a restatement of the claim with $(m-1)$ variables, where by (10), each term $(1 - H(N_0\oplus N_i))$ can be viewed as $I(N_1\oplus N_0\,;\, N_1\oplus N_i)$. This completes the proof. ∎

The above result can be expressed in the following alternative form:

Proposition 5: Let $X, N_0, N_1, \ldots, N_m$ be independent Bernoulli variables. Then the following inequality holds:

$$I(X\oplus N_0\,;\, X\oplus N_1, \ldots, X\oplus N_m) + \sum_{i=1}^{m} H(N_0\oplus N_i) \leq m \qquad (12)$$

Note that X is no longer restricted to have entropy 1.

Proof: The proof follows using the manipulation used in the proof of Proposition 4. Using the expression in (11) for D, we showed that for arbitrary independent Bernoulli variables $N_0, N_1, \ldots, N_m$,

$$I(N_0\oplus N_1\,;\, N_1\oplus N_2, \ldots, N_1\oplus N_m) - \sum_{i=2}^{m} \big(1 - H(N_0\oplus N_i)\big) \leq 0$$

The result follows immediately from the instance of the above expression with $m+2$ Bernoulli variables, and after renaming variable. ∎

Note that using simple upper bounds of 1 on each of the terms in (12), it is trivially seen that the LHS is upper bounded by $m+1$; the stronger result shown here has been somewhat harder to prove.

A more general form of Proposition 4 can be derived for independent BIMSCs:

Proposition 6: Let X be a Bernoulli($\frac{1}{2}$) random variable, and $\mathcal{C}_0, \mathcal{C}_1, \ldots, \mathcal{C}_m$ be independent BIMSCs. Then the following inequality holds:

$$I(\mathcal{C}_0(X)\ ;\ \mathcal{C}_1(X), \ldots, \mathcal{C}_m(X)) \leq \sum_{i=1}^{m} I(\mathcal{C}_0(X)\ ;\ \mathcal{C}_i(X)) \tag{13}$$

Proof: As argued before, the vector channel created from finitely many BIMSCs is itself a BIMSC. It thus suffices to prove the proposition for the case of $m = 2$, and the general statement will follow by simple induction.

We proceed with the notation used in the proof of Proposition 3, to show that

$$\begin{aligned}
&I(\mathcal{C}_0(X)\ ;\ \mathcal{C}_1(X), \mathcal{C}_2(X)) \\
&\quad = \frac{1}{8} \sum_{x,y,z} \pi_0(x)\pi_1(y)\pi_2(z) I(X \oplus N_x\ ;\ X \oplus N_y, X \oplus N_z) \\
&\quad \leq \frac{1}{8} \sum_{x,y,z} \pi_0(x)\pi_1(y)\pi_2(z) \big(\ I(X \oplus N_x\ ;\ X \oplus N_y) \\
&\qquad\qquad + I(X \oplus N_x\ ;\ X \oplus N_z)\ \big) \\
&\quad = I(\mathcal{C}_0(X)\ ;\ \mathcal{C}_1(X)) + I(\mathcal{C}_0(X)\ ;\ \mathcal{C}_2(X)),
\end{aligned}$$

where the first equality follows from a similar argument leading to (7); the inequality follows from application of Proposition 4; and the last equality follows from two applications of (7), noting also that $\sum_{x \in \mathcal{A}_i} \pi_i(x) = 2$. As before, N_x, N_y and N_z are dummy independent Bernoulli random variables with parameters $\bar{p}_0(x), \bar{p}_1(y)$ and $\bar{p}_2(z)$ respectively. ∎

III. Connection with Channel Coding

Consider a Bernoulli($\frac{1}{2}$) random variable X, which is repeatedly observed over binary symmetric channels with noise variables $N_1, \ldots, N_m$.

Then unless X is completely determined by the previous observations, one might want to observe X yet again to reduce uncertainty. We are then interested in determining the amount of redundant information if we were to observe X one more time over another BSC with independent noise N_0. Intuitively, this is the amount of potential information that the observer gives up in using the channel to observe a less-than-completely-random variable, see e.g. [2] and [3] for the related analysis.

The redundant information in the case above is then

$$\mathcal{L} := I(X \oplus N_0\ ;\ X \oplus N_1, \ldots, X \oplus N_m).$$

In practice, this exact expression is not tractable, and it generally suffices to find simple upper bounds on $\mathcal{L}$. Proposition 4 immediately addresses this problem, where $\mathcal{L}$ is bounded in terms of $I(X \oplus N_0\ ;\ X \oplus N_i)$'s, i.e. the amount of redundant information when considering each previous observation individually.

A more realistic setting would involve *indirect observations* of each output variable X, in terms of linear combinations of observations of other output variables that sum up to X. In that case, the effective noise BSC variables $N_1, \ldots, N_m$ are no longer independent, but rather are BSBVs over a common set of base random variables.

IV. Discussion

The results of (8) and (12) are at some level intuitive. In particular, the inequality in (8) resembles some kind of a independence bound on the mutual informations. It may then be expected that a generalization of such result is true, where the variables are not restricted to be all mutually independent. There are however serious pitfalls in expecting too much to be true on these matters. For example, the symmetry assumption that results from the setup of the Proposition 4 seems to be crucial.

As another example, consider the simplest non-trivial case of $m = 1$ in (12): After simple manipulation, it is shown that for independent Bernoulli variables X, Y, Z,

$$\begin{aligned}
H(X \oplus Y) + H(Y \oplus Z) + H(Z \oplus X) \\
- H(X \oplus Y, Y \oplus Z, Z \oplus X) \leq 1
\end{aligned}$$

One might be tempted to try and prove that for arbitrary Bernoulli U, V, W,

$$H(U) + H(V) + H(W) - H(U, V, W) \overset{?}{\leq} 1,$$

but this inequality is certainly not true in general; for example if $U = V = W$ is the same Bernoulli($\frac{1}{2}$) random variable, then the LHS will equal 2.

A slightly less ambitious generalization is the following conjecture for arbitrary Bernoulli variables U and V:

$$H(U) + H(V) + H(U \oplus V) - H(U, V) \overset{?}{\leq} 1. \tag{14}$$

But this conjecture is also not true in general; as a counter-example consider the joint distribution given by $P(U = V = 0) = 0; P(U = 0, V = 1) = P(U = 1, V = 0) = \frac{1}{4}$, and $P(U = V = 1) = \frac{1}{2}$. Then

$$H(U) + H(V) + H(U \oplus V) - H(U, V) = 1 + (1 - \frac{3}{2}\log\frac{3}{2}),$$

which can be readily seen to be larger than 1 from the convexity of the function $\eta(x) := x \log x$, and the fact that $\eta(1) = 0$ and $\eta(2) = 2$.

Indeed, using numerical methods, it appears that the set of joint distributions that satisfy (14), i.e. $\{P_{U,V}(u, v)\ :\ H(U) +$

$H(V) + H(U \oplus V) - H(U, V) \leq 1\}$, is closely approximated by distributions of the form $P_{X \oplus Y, Y \oplus Z}(u, v)$ for independent Bernoulli variables X, Y and Z.

The monotonicity results in Lemma 1 and Proposition 3 resemble some of the results on degradation over BIMSCs as discussed in [5]. It would be interesting to explore to what extent a stronger statement with regards to the monotonicity of the mutual information holds while not requiring the independence assumptions we have used.

V. Acknowldegements

The research of the second author was supported by the National Science Foundation grants CCF-0500023 and CCF-0635372, by Marvell Semiconductor, and by the University of California MICRO program.

References

[1] R. W. Yeung, "A Framework for Linear Information Inequalities," *IEEE Transactions on Information Theory*, 43(11), pp. 1924-1934, 1997.

[2] A. Shokrollahi, "Raptor Codes", *IEEE Transactions on Information Theory*, 52(6), pp. 2551-2567, 2006.

[3] P. Pakzad and A. Shokrollahi, "An Information-Theoretic View on the Analysis of Fountain Codes," in preparation.

[4] T. Cover and J. Thomas, "Elements of Information Theory," *Wiley-Interscience*, New York, 1991.

[5] T. Richardson and R. Urbanke, "Modern Coding Theory," *Cambridge University Press*, New York, 2008.