Lower Bounds on Mutual Information for Linear Codes Transmitted over Binary Input Channels, and for Information Combining

Uri Erez
Electrical Engineering
Tel Aviv University
Tel Aviv, Israel
Email: uri@eng.tau.ac.il

Or Ordentlich
Computer Science and Engineering
Hebrew University of Jerusalem
Jerusalem, Israel
Email: or.ordentlich@mail.huji.ac.il

Shlomo Shamai (Shitz)
Electrical and Computer Engineering
Technion
Haifa, Israel
Email: sshlomo@ee.technion.ac.il

***Abstract*—It has been known for a long time that the mutual information between the input sequence and output sequence of a binary symmetric channel (BSC) is upper bounded by the mutual information between the same input sequence and the output sequence of a binary erasure channel (BEC) with the same capacity. Recently, Samorodnitsky discovered that one may also lower bound the BSC mutual information in terms of the mutual information between the same input sequence and a more capable BEC. In this paper, we strengthen Samorodnitsky's bound for the special case where the input to the channel is distributed uniformly over a linear code. Furthermore, for a general (not necessarily binary) input distribution P_X and channel $W_{Y|X}$, we derive a new lower bound on the mutual information $I(X;Y^n)$ for n transmissions of $X \sim P_X$ through the channel $W_{Y|X}$.**

I. Introduction

Let $P = P_{Y|X}$ and $Q = Q_{Z|X}$ be two channels with a common input alphabet $\mathcal{X}$ and output alphabets $\mathcal{Y}, \mathcal{Z}$, respectively. Three common criteria for comparing/partially-ordering them are [1]–[3]:

- We say that P is degraded with respect to Q if there exists a third channel $W = W_{Y|Z}$ with input alphabet $\mathcal{Z}$ and output alphabet $\mathcal{Y}$ such that $P = W \circ Q$, that is, P is the composition of W and Q.
- We say that P is more noisy than Q if for any distribution P_{UX} we have that $I(U;Y) \leq I(U;Z)$, where Y is obtained by feeding X to P, and Z is obtained by feeding X to Q.
- We say that P is less capable than Q if for any distribution P_X we have that $I(X;Y) \leq I(X;Z)$, where Y is obtained by feeding X to P, and Z is obtained by feeding X to Q.

Clearly if P is degraded with respect Q, it is also more noisy than it, and similarly, if P is more noisy than Q, it is also less capable than it. Furthermore, we have the following tensorization property for the three criteria: If P is degraded with respect to (respectively, more noisy, less capable than) Q, then $P^{\otimes n}$ is degraded with respect to (respectively, more noisy, less capable than) $Q^{\otimes n}$ [1, Problem 6.18], [4], [5]. Here, $P^{\otimes n}$ denotes the product channel from $\mathcal{X}^n$ to $\mathcal{Y}^n$ obtained by applying P independently on each coordinate of X^n.

The above criteria are useful whenever computing mutual information expressions involving P is hard, whereas computing the same expressions with a channel Q which dominates P under the criterion relevant to the problem, is significantly easier. A canonical choice for Q is the erasure channel, which outputs $Z = X$ with probability $1-e$ and outputs $Z =?$ with probability e. This is a convenient choice because computing mutual information expressions involving the erasure channel is often a feasible task, and furthermore, finding the "dirtiest" erasure channel (that is, the erasure channel with the largest e) that is less noisy than P is equivalent to computing the strong data processing inequality (SDPI) coefficient of the channel P [4]. The SDPI coefficient of a channel $P_{Y|X}$ is defined as

$$\eta(P_{Y|X}) = \sup_{P_{UX}} \frac{I(U;Y)}{I(U;X)},$$

where the supremum is with respect to all Markov triplets $U - X \overset{P_{Y|X}}{-} Y$.

Computation of $\eta(P_{Y|X})$ reduces to computation of the SDPI coefficients of all binary sub-channels induced by $P_{Y|X}$, i.e., all channels obtained by restricting the input to two symbols $\{x_0, x_1\} \subset \mathcal{X}$ [6], for which closed-form expressions and bounds exist [4].

A special case that has received considerable attention in the literature is taking the channel $P_{Y|X}$ as a binary symmetric channel (BSC) with capacity $t \in [0,1]$. Let

$$h(p) = -p\log p - (1-p)\log(1-p),$$

and let h^{-1} be its inverse restricted to $[0, 1/2]$, where throughout the paper all logarithms are taken to base 2. It is well known [2, Example 5.4] that this channel is degraded with respect to a binary erasure channel (BEC) with capacity $1 - 2h^{-1}(1-t)$, is more noisy than a BEC channel with capacity $(1 - 2h^{-1}(1-t))^2$, and is less capable than a BEC with capacity t. In particular, for any input distribution X^n and $0 \leq t \leq t_1 \leq 1$, we have that

$$I^{(t)}_{\mathrm{BSC}}(X^n;Y^n) \leq I^{(t_1)}_{\mathrm{BEC}}(X^n;Y^n),$$

where $I^{(t)}_{\mathrm{BSC}}(X^n;Y^n)$ denotes the mutual information between X^n and the output of a memoryless BSC channel with capacity t, and $I^{(t_1)}_{\mathrm{BEC}}(X^n;Y^n)$ the mutual information between X^n and the output of a memoryless BEC channel with capacity t_1.

Thus, in cases where the mutual information between the input vector and the output of a BEC channel can be computed/estimated, we immediately obtain upper bounds on the mutual information for the case of a BSC channel.

In fact, it is well-known that among all binary-input memoryless output-symmetric (BMS) channels (see Definition 1 below) with the same capacity, the BEC is the most capable, and the BSC is the least capable [7]. This implies that for any input distribution X^n, any BMS channel, and $0 \le t_- \le t \le t_+ \le 1$ we have

$$I^{(t_-)}_{\mathrm{BSC}}(X^n;Y^n) \le I^{(t)}_{\mathrm{BMS}}(X^n;Y^n) \le I^{(t_+)}_{\mathrm{BEC}}(X^n;Y^n).$$

Thus, obtaining bounds in the other direction, i.e., lower bounds on $I^{(t)}_{\mathrm{BSC}}(X^n;Y^n)$ in terms of $I^{(t_0)}_{\mathrm{BEC}}(X^n;Y^n)$ is desirable (for some t_0), as it enables to bound $I^{(t)}_{\mathrm{BMS}}(X^n;Y^n)$ from above and below using only mutual information expressions involving the BEC.

However, deriving such bounds is a more challenging task. The main reason for this is that for $t \in (0,1)$, there is no $0 < t_0$ for which the BEC with capacity t_0 is less capable than the BSC with capacity t.

While lower bounds of the form $I^{(t)}_{\mathrm{BSC}}(X^n;Y^n) \ge I^{(t_0)}_{\mathrm{BEC}}(X^n;Y^n)$ that hold for any X^n are impossible to obtain, in [8] Samorodnitsky had the somewhat counter-intuitive observation that we can nevertheless lower bound $I^{(t)}_{\mathrm{BSC}}(X^n;Y^n)$ using the mutual information between X^n and the output of a *less* noisy BEC. In particular, he has shown that for any X^n and any $t_1 \ge \eta_t = (1-2h^{-1}(1-t))^2$, it holds that

$$I^{(t)}_{\mathrm{BSC}}(X^n;Y^n) \ge n \cdot \psi_t\left(\frac{I^{(t_1)}_{\mathrm{BEC}}(X^n;Y^n)}{nt_1}\right), \qquad (1)$$

where $\psi_t : [0,1] \to [0,t]$ is some increasing strictly convex function, to be explicitly specified later.[1]

One of the most interesting applications of Samorodnitsky's result is for the case where X^n is uniformly distributed on some linear code. In this case, the result implies that a code will attain high mutual information when transmitted over the BSC channel, if it attains high mutual information when transmitted over a BEC channel (though with a different capacity). Our main result in this paper is an improvement of Samorodnitsky's result for this special case where $X^n \sim \mathrm{Unif}(C)$ and $C \subset \{0,1\}^n$ is a linear code. For this special case, we improve Samorodnitsky's bound from (1) to

$$I^{(t)}_{\mathrm{BSC}}(X^n;Y^n) \ge n \cdot \bar{\psi}_t\left(\frac{I^{(t_1)}_{\mathrm{BEC}}(X^n;Y^n)}{nt_1}\right),$$

where $\bar{\psi}_t(x) : [0,1] \to [0,t] = t \cdot x$ is the upper concave envelope of $\psi_t(x)$.

[1] In fact, [8] establishes this only for $t_1 = \eta_t$, but by Lemma 1 proved in the appendix, this holds for all $t_1 \ge \eta_t$.

An important special case of a linear code is the repetition code. A uniform distribution on this code corresponds to $X^n = (X, \cdots, X)$ where $X \sim \mathrm{Bern}(1/2)$. For this case the problem of comparing $I^{(t)}_{\mathrm{BSC}}(X^n;Y^n) = I^{(t)}_{\mathrm{BSC}}(X;Y^n)$ to $I^{(t_1)}_{\mathrm{BEC}}(X^n;Y^n) = I^{(t_1)}_{\mathrm{BEC}}(X;Y^n)$ is referred to as the *information combining* problem, which has been studied extensively in the literature [9], [10]. While the case of $X \sim \mathrm{Bern}(1/2)$ transmitted through n copies of a BSC channel is handled by our main result, we further derive a lower bound for the general case where $X \sim P_X$ is transmitted n times through a channel $W^{\otimes n}$, and show that

$$I(X;Y^n) \ge \frac{I(P_X;W)}{\eta(P_X,W)}(1-(1-\eta(P_X,W))^n), \qquad (2)$$

where Y^n is the output of the channel when X is transmitted n times, $I(P_X,W) = I(X;Y_1)$, and

$$\eta(P_X,W) = \sup_{P_{U|X}} \frac{I(U;Y)}{I(U;X)} \qquad (3)$$

is the input-dependent SDPI coefficient of the channel W with input P_X. Note that we can further lower bound (2) as

$$I(X;Y^n) \ge \frac{1-e^{-n\cdot\eta(P_X,W)}}{\eta(P_X,W)} \cdot I(P_X,W), \qquad (4)$$

which is close to the obvious upper bound $nI(P_X;W)$ for $n \cdot \eta(P_X,W) \ll 1$. Thus, our bound essentially shows that when $n \cdot \eta(P_X,W) \ll 1$ each measurement contributes about $I(P_X,W)$ bits of information to $I(X;Y^n)$ (as is the case for i.i.d. transmission).

II. Main Result

For a random vector X^n on $\{0,1\}^n$, we denote by $I^{(t)}_{\mathrm{BEC}}(X^n;Y^n)$, respectively $I^{(t)}_{\mathrm{BSC}}(X^n;Y^n)$, the mutual information between X^n and the output of a memoryless BEC channel, respectively BSC channel, with capacity t. We denote the strong data processing inequality (SDPI) coefficient of a BSC channel with capacity t by [11]

$$\eta_t = (1-2h^{-1}(1-t))^2. \qquad (5)$$

We further denote the ratio between the capacity and the SDPI coefficient by

$$\alpha_t = \frac{t}{\eta_t} = \frac{t}{(1-2h^{-1}(1-t))^2}. \qquad (6)$$

It can be verified that for all $0 \le t \le 1$ we have $t \le \eta_t$, and consequently, $\alpha_t \le 1$. Furthermore, for all $0 < t \le 1$,

$$\alpha_t > \frac{\log_2(e)}{2}.$$

Our main result is the following.

Theorem 1: Let $C \subset \{0,1\}^n$ be a linear code, $u \in \{0,1\}^n$ be some shift, and $X^n = X^n_{C,u} \sim \mathrm{Uniform}(C+u)$. Then

$$I^{(t)}_{\mathrm{BSC}}(X^n;Y^n) \ge t \cdot \frac{I^{(\eta_t)}_{\mathrm{BEC}}(X^n;Y^n)}{\eta_t} = \alpha_t \cdot I^{(\eta_t)}_{\mathrm{BEC}}(X^n;Y^n). \qquad (7)$$

Remark 1: We may rewrite (7) as

$$\frac{I^{(t)}_{\mathrm{BSC}}(X^n;Y^n)}{nt} \geq \frac{I^{(\eta_t)}_{\mathrm{BEC}}(X^n;Y^n)}{n\eta_t}, \tag{8}$$

indicating that for (shifted) linear codes, the fraction of capacity over the BSC with capacity t is at least as large as the fraction of capacity over a BEC with capacity η_t.

Remark 2: In [8, Theorem 12] (see also [4] and [12]), Samorodnitsky proved that for the BSC with capacity t, for any input X^n on $\{0,1\}^n$ it holds that

$$H(Y^n) \geq n \cdot \varphi_t\left(\frac{I^{(\eta_t)}_{\mathrm{BEC}}(X^n;Y^n)}{n\cdot\eta_t}\right), \tag{9}$$

where $\varphi_t(x) = h\left(h^{-1}(1-t)\star h^{-1}(x)\right)$ is the function from Mrs. Gerber's Lemma [13]. Here, $a\star b = a(1-b)+b(1-a)$ is the convolution between two numbers $a,b\in[0,1]$. Subtracting $n(1-t)$ from both sides of (9), we obtain

$$\begin{aligned} I^{(t)}_{\mathrm{BSC}}(X^n;Y^n) &= H(Y^n) - H(Y^n|X^n) = H(Y^n) - n(1-t) \\ &\geq n\cdot\varphi_t\left(\frac{I^{(\eta_t)}_{\mathrm{BEC}}(X^n;Y^n)}{n\cdot\eta_t}\right) - n(1-t) \\ &= n\cdot\psi_t\left(\frac{I^{(\eta_t)}_{\mathrm{BEC}}(X^n;Y^n)}{n\cdot\eta_t}\right), \end{aligned} \tag{10}$$

where $\psi_t(x) = \varphi_t(x) - (1-t)$ is defined for $0\leq x\leq 1$. Since $x\mapsto\varphi_t(x)$ is convex, so is $x\mapsto\psi_t(x)$. Noting further that $\psi_t(0)=0$ and $\psi_t(1)=t$, convexity implies that $\psi_t(x)\leq t\cdot x$. In particular,

$$\psi_t\left(\frac{I^{(\eta_t)}_{\mathrm{BEC}}(X^n;Y^n)}{n\cdot\eta_t}\right) \leq t\cdot\frac{I^{(\eta_t)}_{\mathrm{BEC}}(X^n;Y^n)}{n\cdot\eta_t}. \tag{11}$$

Comparing this with Theorem 1, we see that our bound is always at least as good as Samorodnitsky's. However, while Samorodnitsky's lower bound on $I^{(t)}_{\mathrm{BSC}}(X^n;Y^n)$ is valid for *any* input X^n, our bound is only valid for X^n uniform on a shifted linear code. In fact, it is easy to verify that Theorem 1 does not hold if one does not impose any assumptions on X^n. Indeed, by the convexity of $t\mapsto g(t) = h(p\star h^{-1}(1-t)) - (1-t)$, and the fact that $g(0)=0$ and $g(1)=h(p)$, it follows that $g(t)\leq t h(p)$. Thus, for $X^n\sim\mathrm{Bern}^{\otimes n}(p)$ we have

$$\frac{I^{(t)}_{\mathrm{BSC}}(X^n;Y^n)}{t} = \frac{ng(t)}{t} \leq nh(p) = \frac{I^{(\eta_t)}_{\mathrm{BEC}}(X^n;Y^n)}{\eta_t}. \tag{12}$$

Definition 1 (BMS channels): A memoryless channel with binary input X and output Y is called *binary-input memoryless output-symmetric (BMS)* if there exists a sufficient statistic $T(Y) = (X\oplus Z_A, A)$ for X, where (A, Z_A) are statistically independent of X, and Z_A is a binary random variable with $\Pr(Z_A = 1|A=a) = a$.

It is well known and easy to verify that among all BMS channels with capacity t, the BEC is the most capable one, whereas the BSC is the least capable, see e.g. [7]. Thus, the following is a straightforward corollary of Theorem 1.

Corollary 1: Under the assumptions of Theorem 1, for any BMS channel with capacity t we have

$$\alpha_t\cdot I^{(\eta_t)}_{\mathrm{BEC}}(X^n;Y^n) \leq I^{(t)}_{\mathrm{BMS}}(X^n;Y^n) \leq I^{(t)}_{\mathrm{BEC}}(X^n;Y^n).$$

Furthermore, since $t\leq\eta_t$, we have that the BEC with capacity t is degraded with respect to the BEC with capacity η_t. Thus, the following statement immediately follows from Corollary 1.

Corollary 2: Under the assumptions of Theorem 1,

$$\alpha_t\cdot I^{(t)}_{\mathrm{BEC}}(X^n;Y^n) \leq I^{(t)}_{\mathrm{BMS}}(X^n;Y^n) \leq I^{(t)}_{\mathrm{BEC}}(X^n;Y^n)$$

where α_t, defined in (6), satisfies $\alpha_t > \frac{\log_2(e)}{2}$, for all $0<t\leq 1$.

Proof of Theorem 1. We may assume without loss of generality that $\mathrm{rank}(C)>0$, since otherwise X^n is deterministic, so that $I^{(\eta_t)}_{\mathrm{BEC}}(X^n;Y^n) = I^{(t)}_{\mathrm{BSC}}(X^n;Y^n) = 0$ and the statement holds trivially.

Since X^n is uniform over a shifted linear code with positive rank, we have that $X_i\sim\mathrm{Bern}(1/2)$ for all $i=1,\ldots,n$, and in particular

$$I^{(\eta_t)}_{\mathrm{BEC}}(X_i;Y_i) = \eta_t,\ I^{(t)}_{\mathrm{BSC}}(X_i;Y_i) = t, \quad \forall i=1,\ldots,n. \tag{13}$$

This also implies the statement for $n=1$. We proceed by induction. Assume the statement holds for all linear codes and shifts in $\{0,1\}^{n-1}$.

Note that for any memoryless channel, and in particular for the BEC and the BSC, we have that

$$\begin{aligned} I(X^n;Y^n) &= I(X^{n-1},X_n;Y^{n-1},Y_n) \\ &= I(X^{n-1};Y^{n-1},Y_n) + I(X_n;Y^{n-1},Y_n|X^{n-1}) \\ &= I(X^{n-1};Y^{n-1}) + I(X^{n-1};Y_n|Y^{n-1}) + I(X_n;Y_n|X^{n-1}) \\ &= I(X^{n-1};Y^{n-1}) + I(X_n;Y_n) - I(Y^{n-1};Y_n). \end{aligned} \tag{14}$$

By (13), for the BEC, we therefore have that

$$\begin{aligned} I^{(\eta_t)}_{\mathrm{BEC}}(X^n;Y^n) &= I^{(\eta_t)}_{\mathrm{BEC}}(X^{n-1};Y^{n-1}) + \eta_t - I^{(\eta_t)}_{\mathrm{BEC}}(Y^{n-1};Y_n) \\ &= I^{(\eta_t)}_{\mathrm{BEC}}(X^{n-1};Y^{n-1}) + \eta_t - \eta_t I^{(\eta_t)}_{\mathrm{BEC}}(Y^{n-1};X_n), \end{aligned} \tag{15}$$

while for the BSC, we have that

$$\begin{aligned} I^{(t)}_{\mathrm{BSC}}(X^n;Y^n) &= I^{(t)}_{\mathrm{BSC}}(X^{n-1};Y^{n-1}) + t - I^{(t)}_{\mathrm{BSC}}(Y^{n-1};Y_n) \\ &\geq I^{(t)}_{\mathrm{BSC}}(X^{n-1};Y^{n-1}) + t - \eta_t I^{(t)}_{\mathrm{BSC}}(Y^{n-1};X_n). \end{aligned} \tag{16}$$

In the last inequality we have used the strong data processing inequality (SDPI), stating that for any $U - X_n - Y_n$, where $P_{Y_n|X_n}$ is a BSC of capacity t, we have that $I(U;Y_n)\leq\eta_t I(U;X_n)$. Since $Y^{n-1} - X_n - Y_n$ forms a Markov chain in this order, we can indeed apply the SDPI with $U = Y^{n-1}$ and obtain $I^{(t)}_{\mathrm{BSC}}(Y^{n-1};Y_n)\leq\eta_t I^{(t)}_{\mathrm{BSC}}(Y^{n-1};X_n)$. We continue by noting that, since $X_n - X^{n-1} - Y^{n-1}$ forms a Markov chain in this order, we have

$$\begin{aligned} &I^{(t)}_{\mathrm{BSC}}(Y^{n-1};X_n) \\ &\quad = I^{(t)}_{\mathrm{BSC}}(Y^{n-1};X^{n-1}) - I^{(t)}_{\mathrm{BSC}}(Y^{n-1};X^{n-1}|X_n). \end{aligned} \tag{17}$$

Substituting (17) into (16), gives

$$I^{(t)}_{\mathrm{BSC}}(X^n;Y^n) \geq (1-\eta_t) I^{(t)}_{\mathrm{BSC}}(X^{n-1};Y^{n-1}) + t + \eta_t I^{(t)}_{\mathrm{BSC}}(Y^{n-1};X^{n-1}|X_n). \quad (18)$$

The random variable X^{n-1} is uniformly distributed over the projection of $C+u$ to the first $n-1$ coordinates. Since this projection is a shifted linear code in $\{0,1\}^{n-1}$, by the induction hypothesis, we have

$$I^{(t)}_{\mathrm{BSC}}(X^{n-1};Y^{n-1}) \geq \alpha_t I^{(\eta_t)}_{\mathrm{BEC}}(X^{n-1};Y^{n-1}). \quad (19)$$

Furthermore, conditioned on $X_n=0$ or $X_n=1$, we also have that X^{n-1} is uniformly distributed over a shifted linear code in $\{0,1\}^{n-1}$ (though those shifted linear codes may differ for $X_n=0$ and $X_n=1$). Thus, again by the induction hypothesis

$$\begin{aligned}
&I^{(t)}_{\mathrm{BSC}}(X^{n-1};Y^{n-1}|X_n) \\
&= \frac{1}{2} I^{(t)}_{\mathrm{BSC}}(X^{n-1};Y^{n-1}|X_n=0) \\
&+ \frac{1}{2} I^{(t)}_{\mathrm{BSC}}(X^{n-1};Y^{n-1}|X_n=1) \\
&\geq \frac{\alpha_t}{2} I^{(\eta_t)}_{\mathrm{BEC}}(X^{n-1};Y^{n-1}|X_n=0) \\
&+ \frac{\alpha_t}{2} I^{(\eta_t)}_{\mathrm{BEC}}(X^{n-1};Y^{n-1}|X_n=1) \\
&= \alpha_t I^{(\eta_t)}_{\mathrm{BEC}}(X^{n-1};Y^{n-1}|X_n) \\
&= \alpha_t \left(I^{(\eta_t)}_{\mathrm{BEC}}(X^{n-1};Y^{n-1}) - I^{(\eta_t)}_{\mathrm{BEC}}(X_n;Y^{n-1}) \right) \quad (20)
\end{aligned}$$

where the last equality holds since $Y^{n-1}-X^{n-1}-X_n$ forms a Markov chain in this order, as in (17). Substituting (19) and (20) into (18), we obtain

$$\begin{aligned}
&I^{(t)}_{\mathrm{BSC}}(X^n;Y^n) \geq \alpha_t(1-\eta_t) I^{(\eta_t)}_{\mathrm{BEC}}(X^{n-1};Y^{n-1}) + t \\
&+ \alpha_t \eta_t \left(I^{(\eta_t)}_{\mathrm{BEC}}(X^{n-1};Y^{n-1}) - I^{(\eta_t)}_{\mathrm{BEC}}(X_n;Y^{n-1}) \right) \\
&= \alpha_t I^{(\eta_t)}_{\mathrm{BEC}}(X^{n-1};Y^{n-1}) + t - \alpha_t \eta_t I^{(\eta_t)}_{\mathrm{BEC}}(X_n;Y^{n-1}) \\
&= \alpha_t \left[I^{(\eta_t)}_{\mathrm{BEC}}(X^{n-1};Y^{n-1}) + \eta_t - \eta_t I^{(\eta_t)}_{\mathrm{BEC}}(X_n;Y^{n-1}) \right] \\
&= \alpha_t I^{(\eta_t)}_{\mathrm{BEC}}(X^n;Y^n), \quad (21)
\end{aligned}$$

where in the last equality we have used (15) and the fact that $\alpha_t = \frac{t}{\eta_t}$. This completes the proof. ■

III. Information Combining

Let $X \sim P_X$, and let $W = W_{Y|X}$ be some channel with input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$. Assume X is transmitted n times through W, and the output is $Y^n = (Y_1,\ldots,Y_n)$. What can we say about $I(X;Y^n)$? Since the channel from $X^n=(X,\ldots,X)$ to Y^n is memoryless, we have that

$$I(X;Y^n) \leq \sum_{i=1}^{n} I(X;Y_i) = nI(X;Y) = nI(P_X,W). \quad (22)$$

Combining this with the trivial upper bound $I(X;Y^n) \leq H(X) = H(P_X)$, we have that $I(X;Y^n) \leq \min\{H(P_X), nI(P_X,W)\}$. Denote by EC_e the erasure channel with input $\mathcal{X}$ whose output is X with probability $1-e$ and ? with probability e. Let

$$\begin{aligned}
C_{\mathrm{MC}} &= C_{\mathrm{MC}}(W) \\
&= \min\{1-e \in [0,1] \ : \ W \text{ is less capable than } \mathrm{EC}_e\}.
\end{aligned}$$

By tensorization of the more capable partial order, we have

$$\begin{aligned}
I(X;Y^n) = I(P_X, W^{\otimes n}) &\leq I(P_X, \mathrm{EC}^{\otimes n}_{1-C_{\mathrm{MC}}}) \\
&= (1-(1-C_{\mathrm{MC}})^n) H(P_X). \quad (23)
\end{aligned}$$

Our main result is a lower bound on $I(X;Y^n)$ taking a similar form to (23) .

Theorem 2: Let $X \sim P_X$ and $W = W_{Y|X}$ be an input distribution and a channel with input-dependant SDPI coefficient satisfying $\eta(P_X,W) \leq \eta$. Assume X is transmitted n times through W, and the output is $Y^n=(Y_1,\ldots,Y_n)$. Then,

$$I(X;Y^n) \geq \alpha(1-(1-\eta)^n) H(P_X), \quad (24)$$

where

$$\alpha = \frac{I(P_X,W)}{\eta H(P_X)}. \quad (25)$$

Proof. For $n=1$ the claim holds with equality. We proceed by induction. Starting from (14), we have

$$\begin{aligned}
I(X;Y^n) &= I(X;Y^{n-1}) + I(P_X,W) - I(Y^{n-1};Y_n) \\
&\geq I(X;Y^{n-1}) + I(P_X,W) - \eta I(Y^{n-1};X) \quad (26) \\
&= (1-\eta) I(X;Y^{n-1}) + I(P_X,W), \quad (27)
\end{aligned}$$

where (26) follows from the strong data processing inequality, as $Y^{n-1}-X-Y_n$ forms a Markov chain in this order. Using the induction hypothesis $I(X;Y^{n-1}) \geq \alpha(1-(1-\eta)^{n-1}) H(P_X)$, we further lower bound (27) as

$$\begin{aligned}
I(X;Y^n) &\geq \alpha(1-(1-\eta)^{n-1})(1-\eta) H(P_X) + I(P_X,W) \\
&= \alpha \left[(1-(1-\eta)^{n-1})(1-\eta) H(P_X) + \frac{I(P_X;W)}{\alpha} \right] \\
&= \alpha \left[(1-(1-\eta)^{n-1})(1-\eta) H(P_X) + \eta H(P_X) \right] \\
&= \alpha(1-(1-\eta)^n) H(P_X), \quad (28)
\end{aligned}$$

which establishes the claim. ■

Remark 3: Recall that the information bottleneck curve corresponding to $(X,Y) \sim P_{XY}$ is defined as

$$\mathrm{IB}_{P_{XY}}(R) = \max\left\{ I(U;Y) \ : \ I(U;X) \leq R, U-X-Y \right\}.$$

Since $R \mapsto \mathrm{IB}_{P_{XY}}(R) \in [0, H(P_X)]$ is concave [14], [15], and satisfies $\mathrm{IB}_{P_{XY}}(0)=0$ and $\mathrm{IB}_{P_{XY}}(H(P_X)) = I(P_X;W)$, we have that $\mathrm{IB}_{P_{XY}}(R) \geq \frac{I(P_X;W)}{H(P_X)} R$. This implies that[2]

$$\eta(P_X,W) = \sup_{R \in (0,H(P_X)]} \frac{\mathrm{IB}_{P_{XY}}(R)}{R} \geq \frac{I(P_X;W)}{H(P_X)}. \quad (29)$$

[2]In fact, the supremum in (29) is attained for $R \to 0$ [16], [17]

This, in turn, shows that $\alpha \leq 1$.

Remark 4 (Special case of $P_X = \mathrm{Bern}(1/2)$ and $W = \mathrm{BSC}$): Note that for the special case of $\mathcal{X} = \{0,1\}$, $P_X = \mathrm{Bern}(1/2)$, and W taken as a BSC with capacity t the conclusion of Theorem 2 follows from Theorem 1. This follows since in this case $X^n = (X, \cdots, X)$ can be viewed as a random codeword in the repetition code (which is of course linear), and furthermore, for this choice of W and P_X we have [11] that $\eta(P_X, W) = \eta(W) = (1-2h^{-1}(t))^2$. We may further relax Theorem 2, as in Corollary 2, lower bounding $\eta_t = (1-2h^{-1}(1-t))^2$ by t, to obtain that for $X \sim \mathrm{Bern}(1/2)$ we have

$$\alpha_t \cdot I^{(t)}_{\mathrm{BEC}}(X;Y^n) \leq I^{(t)}_{\mathrm{BSC}}(X;Y^n) \leq I^{(t)}_{\mathrm{BEC}}(X;Y^n). \quad (30)$$

Recalling that $\alpha_t \geq \frac{\log(e)}{2}$ for all $0 < t \leq 1$, we obtain the uniform bound

$$I^{(t)}_{\mathrm{BSC}}(X;Y^n) \geq \frac{\log(e)}{2} I^{(t)}_{\mathrm{BEC}}(X;Y^n) = \frac{\log(e)}{2}(1-(1-t)^n). \quad (31)$$

Equation (31) may be tightened. Specifically, numerical evidence suggests that $I(X;Y^n) > 0.92(1-(1-t)^n)$, which is significantly tighter than the uniform lower bound (31) (recall that $\frac{\log(e)}{2} \approx 0.72$). To pursue such improvements, one may tighten the inequality in (26) by bounding $I(Y^{n-1};Y_n) \leq \mathrm{IB}_{P_{XY}}(I(Y^{n-1};X)$ which has a closed-form solution [18] for $P_X = \mathrm{Bern}(1/2)$ and $W = \mathrm{BSC}$.

IV. Applications

Definition 2: We say that a code $C \subset \{0,1\}^n$ of rate R is ε-information-capacity achieving for the BEC, if for any $t > R$ we have that $I^{(t)}_{\mathrm{BEC}}(X^n;Y^n) \geq n(R-\varepsilon)$, where $X^n \sim \mathrm{Uniform}(C)$.

We show that for codes that are ε-information-capacity achieving for the BEC, the corresponding mutual information over the BSC cannot be too small. Results in similar spirit have been obtained in [19], see also [20].

Theorem 3: Let $C \subset \{0,1\}^n$ be a linear code of rate R, that is ε-information-capacity achieving for the BEC. Then,

$$I^{(t)}_{\mathrm{BSC}}(X^n;Y^n) \geq n\left(1-\frac{\varepsilon}{R}\right) \cdot \begin{cases} t & t < 1-h\left(\frac{1-\sqrt{R}}{2}\right) \\ \frac{tR}{\eta_t} & t \geq 1-h\left(\frac{1-\sqrt{R}}{2}\right) \end{cases}. \quad (32)$$

Proof. Let $t^* = 1-h\left(\frac{1-\sqrt{R}}{2}\right)$, be such that $\eta_{t^*} = R$. Then,

$$\frac{I^{(\eta_{t^*})}_{\mathrm{BEC}}(X^n;Y^n)}{\eta_{t^*}} \geq \frac{n(R-\varepsilon)}{R} = n\left(1-\frac{\varepsilon}{R}\right). \quad (33)$$

By Lemma 1 in the Appendix,

$$t \mapsto \frac{I^{(t)}_{\mathrm{BEC}}(X^n;Y^n)}{t} \quad (34)$$

is non increasing, and hence, for all $t \leq t^*$, it holds that $\frac{I^{(\eta_t)}_{\mathrm{BEC}}(X^n;Y^n)}{\eta_t} \geq n\left(1-\frac{\varepsilon}{R}\right)$. By Theorem 1, we therefore have that

$$I^{(t)}_{\mathrm{BSC}}(X^n;Y^n) \geq nt\left(1-\frac{\varepsilon}{R}\right), \quad t \leq t^*. \quad (35)$$

For $t > t^*$ we have that $I^{(t)}_{\mathrm{BEC}}(X^n;Y^n) > n(R-\varepsilon)$, by the assumption that C is a rate R code that is ε-information-capacity achieving for the BEC. Thus, by Theorem 1

$$I^{(t)}_{\mathrm{BSC}}(X^n;Y^n) \geq t \cdot \frac{n(R-\varepsilon)}{\eta_t}, \quad t > t^*, \quad (36)$$

which establishes our claim. ■

Appendix

Lemma 1: Fix an input distribution X^n on $\{0,1\}^n$. The mapping $t \mapsto \frac{I^{(t)}_{\mathrm{BEC}}(X^n;Y^n)}{t}$ is non increasing.

Proof. Let $0 \leq t_1 \leq t \leq 1$. Note that a BEC with capacity t_1 can be obtained by concatenating a BEC with capacity t, denoted $P^{\otimes n}_{Y|X}$ and a $\mathrm{EC}_{1-t'}$, where $t' = \frac{t_1}{t}$, denoted $W^{\otimes n}_{Z|Y}$, where $\mathcal{X} = \{0,1\}$ and $\mathcal{Y} = \mathcal{Z} = \{0,?,1\}$. Let $S_t \subset [n]$ denote the (random) set of indices not erased by $P^{\otimes n}_{Y|X}$ and $S_{t'}$ the (random) set of indices not erased by $W^{\otimes n}_{Y|X}$. For a set $S \subset [n]$ we denote by X_S the restriction of X^n to the indices included in S. We have

$$\begin{aligned} I(X^n;Z^n) &= I(X^n; X_{S_t \cap S_{t'}}, S_t \cap S_{t'}) \\ &= I(X^n; X_{S_t \cap S_{t'}} | S_t \cap S_{t'}) \quad (37) \\ &= H(X_{S_t \cap S_{t'}} | S_t \cap S_{t'}) \\ &= H(X_{S_t \cap S_{t'}} | S_t, S_{t'}), \quad (38) \end{aligned}$$

where (37) follows since $(S_t, S_{t'})$ are statistically independent of X^n. We have

$$\begin{aligned} H(X_{S_t \cap S_{t'}} | S_t, S_{t'}) &= \mathbb{E}_{s_t \sim P_{S_t}}[H(X_{S'_t \cap s_t} | S_{t'}, S_t = s_t)] \\ &\geq t' \cdot \mathbb{E}_{s_t \sim P_{S_t}}[H(X_{s_t} | S_t = s_t)] \quad (39) \\ &= t' \cdot H(X_{S_t} | S_t) \\ &= t' \cdot I(X^n;Y^n), \quad (40) \end{aligned}$$

where in (39) we have used Shearer's Lemma, see e.g. [3, Theorem 1.8]. Recalling that $t' = \frac{t_1}{t}$, we have therefore obtained that

$$\frac{I(X^n;Z^n)}{t_1} \geq \frac{I(X^n;Y^n)}{t}, \quad (41)$$

as claimed. ■

Acknowledgment

The work of Or Ordentlich was supported by the Israel Science Foundation (ISF), grant No.1641/21. The work of Uri Erez was supported by the Israel Science Foundation (ISF), grant No. 588/19 and 736/23. The work of Shlomo Shamai (Shitz) was supported by the Israel Science Foundation (ISF), grant No. 1897/19.

REFERENCES

[1] I. Csiszár and J. Körner, *Information theory: coding theorems for discrete memoryless systems*. Cambridge University Press, 2011.

[2] A. El Gamal and Y.-H. Kim, *Network information theory*. Cambridge University Press, 2011.

[3] Y. Polyanskiy and Y. Wu, "Information theory: From coding to learning," *Book draft*, 2022.

[4] ——, "Strong data-processing inequalities for channels and Bayesian networks," in *Convexity and Concentration*. Springer, 2017, pp. 211–249.

[5] A. Makur and Y. Polyanskiy, "Comparison of channels: Criteria for domination by a symmetric channel," *IEEE Transactions on Information Theory*, vol. 64, no. 8, pp. 5704–5725, 2018.

[6] O. Ordentlich and Y. Polyanskiy, "Strong data processing constant is achieved by binary inputs," *IEEE Transactions on Information Theory*, vol. 68, no. 3, pp. 1480–1481, 2021.

[7] E. Sasoglu, "Polar coding theorems for discrete systems," EPFL, Tech. Rep., 2011.

[8] A. Samorodnitsky, "On the entropy of a noisy function," *IEEE Transactions on Information Theory*, vol. 62, no. 10, pp. 5446–5464, 2016.

[9] I. Land, J. Huber *et al.*, "Information combining," *Foundations and Trends® in Communications and Information Theory*, vol. 3, no. 3, pp. 227–330, 2006.

[10] I. Sutskover, S. Shamai, and J. Ziv, "Extremes of information combining," *IEEE Transactions on Information Theory*, vol. 51, no. 4, pp. 1313–1325, 2005.

[11] R. Ahlswede and P. Gács, "Spreading of sets in product spaces and hypercontraction of the markov operator," *The annals of probability*, pp. 925–939, 1976.

[12] O. Ordentlich, "Novel lower bounds on the entropy rate of binary hidden markov processes," in *2016 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2016, pp. 690–694.

[13] A. Wyner and J. Ziv, "A theorem on the entropy of certain binary sequences and applications–I," *IEEE Transactions on Information Theory*, vol. 19, no. 6, pp. 769–772, 1973.

[14] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," *arXiv preprint physics/0004057*, 2000.

[15] H. Witsenhausen and A. Wyner, "A conditional entropy bound for a pair of discrete random variables," *IEEE Transactions on Information Theory*, vol. 21, no. 5, pp. 493–501, 1975.

[16] V. Anantharam, A. Gohari, S. Kamath, and C. Nair, "On maximal correlation, hypercontractivity, and the data processing inequality studied by Erkip and Cover," *arXiv preprint arXiv:1304.6133*, 2013.

[17] A. Makur, "Information contraction and decomposition," Ph.D. dissertation, Massachusetts Institute of Technology, 2019.

[18] E. Erkip and T. M. Cover, "The efficiency of investment information," *IEEE Transactions on information theory*, vol. 44, no. 3, pp. 1026–1040, 1998.

[19] J. Hązła, A. Samorodnitsky, and O. Sberlo, "On codes decoding a constant fraction of errors on the BSC," in *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, 2021, pp. 1479–1488.

[20] M. Pathegama and A. Barg, "Smoothing of binary codes, uniform distributions, and applications," *Entropy*, vol. 25, no. 11, p. 1515, 2023.