Lautum Information

Daniel P. Palomar, *Member, IEEE*, and Sergio Verdú, *Fellow, IEEE*

***Abstract*—A popular way to measure the degree of dependence between two random objects is by their mutual information, defined as the divergence between the joint and product-of-marginal distributions. We investigate an alternative measure of dependence: the *lautum information* defined as the divergence between the product-of-marginal and joint distributions, i.e., swapping the arguments in the definition of mutual information. Some operational characterizations and properties are provided for this alternative measure of information.**

***Index Terms*—Divergence, hypothesis testing, information measures, Kelly gambling, mutual information.**

I. Introduction

One way to gauge the statistical dependence between two random objects X and Y is the *mutual information* defined as the divergence between the joint and product-of-marginal distributions:[1]

$$I(X;Y) \triangleq D(P_{XY} \,\|\, P_X P_Y) \tag{1}$$

$$= \mathbb{E}\left[\log \frac{P_{XY}(X,Y)}{P_X(X)\, P_Y(Y)}\right]. \tag{2}$$

Since the inception of information theory [1], mutual information has proven to be a key measure of dependence with meaningful operational characterizations, foremost among which is its role in the capacity and rate–distortion function of ergodic channels and sources. Mutual information has also proven a popular measure of statistical dependence in many experimental applications such as neurobiology [2], genetics [3], machine learning [4], [5], medical imaging [6], linguistics [7], artificial intelligence [8], authentication [9], and signal processing [10].

The purpose of this paper is to provide several operational characterizations and a number of useful properties for an alternative measure of dependence where the roles of the joint and product-of-marginal distributions are swapped. We define the *lautum information*[2] between X and Y as

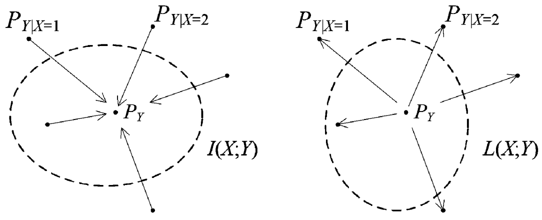

Fig. 1. Geometric visualization of mutual information $I(X;Y)$ and lautum information $L(X;Y)$.

$$L(X;Y) \triangleq D(P_X P_Y \,\|\, P_{XY}) \tag{3}$$

$$= \mathbb{E}\left[\log \frac{P_X(\bar{X})\, P_Y(\bar{Y})}{P_{XY}(\bar{X},\bar{Y})}\right] \tag{4}$$

where the random variables $(\bar{X},\bar{Y})$ are independent with the same marginals as (X,Y).

The difference between the definitions of mutual information and lautum information is illustrated geometrically in Fig. 1 using the alternative expressions

$$I(X;Y) = D\left(P_{Y|X} \,\|\, P_Y \mid P_X\right) \tag{5}$$

$$L(X;Y) = D\left(P_Y \,\|\, P_{Y|X} \mid P_X\right). \tag{6}$$

As usual, general definitions that encompass the general (nondiscrete, noncontinuous) case can be given by defining the Radon–Nikodym derivative

$$Z = \frac{dP_{XY}}{d(P_X \times P_Y)}. \tag{7}$$

Then

$$I(X;Y) = \mathbb{E}[\log Z] \tag{8}$$

$$L(X;Y) = \mathbb{E}\left[\frac{1}{Z}\log\frac{1}{Z}\right] \tag{9}$$

provided Z and the expectations exist, otherwise the corresponding measure is equal to ∞ by convention.

Lautum information does not fall within the class of Shannon-type information measures (i.e., measures that can be expressed as a linear combination of joint entropies [11], [12]). To see that $L(X;Y)$ cannot be written as a linear combination of $H(X)$, $H(Y)$, and $H(X,Y)$, simply note that $L(X;X) = +\infty$ whenever $H(X) > 0$, whereas $H(X)$ is finite for a finite alphabet.

Even before Kullback and Leibler introduced $D(P\|Q)$ [13],[3] Jeffreys [15] introduced the symmetrized form $D(P\|Q) + D(Q\|P)$. Mutual information is, arguably, the most important

Manuscript received June 28, 2006; revised July 31, 2007. This work was supported in part by the Fulbright Program and the Ministry of Education and Science of Spain; the U.S. National Science Foundation under Grant NCR-0074277; and through collaborative participation in the Communications and Networks Consortium sponsored by the U.S. Army Research Laboratory under the Collaborative Technology Alliance Program, Cooperative Agreement DAAD19-01-2-0011. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation thereon. The material in this paper was presented in part at the IEEE Information Theory Workshop (ITW), Punta del Este, Uruguay, March 2006.

D. P. Palomar was with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA. He is now with the Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong (e-mail: palomar@ust.hk).

S. Verdú is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: verdu@princeton.edu).

Communicated by Y. Steinberg, Associate Editor for Shannon Theory.

Digital Object Identifier 10.1109/TIT.2007.915715

[1] P denotes either a probability mass function (pmf) or a probability density function (pdf) unless otherwise indicated.

[2] *Lautum* ("elegant" in Latin) is the reverse spelling of *mutual*.

[3] Originally used, but not defined, in [14].

specialization of divergence, and the corresponding symmetrized form involves the sum of mutual and lautum informations. Both mutual and lautum informations are special cases of Csiszár's f-divergence [16] between the product-of-marginals and joint distributions for different convex functions f.

The paper is organized as follows. Section II describes some operational characterizations of lautum information. Section III explores whether and how the familiar properties of mutual information find counterparts in lautum information. Sections IV and V analyze in detail lautum information for the binary-symmetric channel (BSC) and in the Gaussian case, respectively.

II. Operational Characterizations of Lautum Information

A. Non-Bayesian Testing of Independence

1) Nonasymptotic Setting: Consider a random object X drawn from the distribution $P_{X|H_0}$ under hypothesis H_0 and from $P_{X|H_1}$ under hypothesis H_1. A simple application of the divergence data processing theorem shows that any hypothesis test that achieves

$$\pi_{0|1} = \Pr\left[\text{decide } H_0 \mid H_1 \text{ is true}\right] \tag{10}$$
$$\pi_{1|0} = \Pr\left[\text{decide } H_1 \mid H_0 \text{ is true}\right] \tag{11}$$

must satisfy [17, p. 74][4]

$$d\left(\pi_{0|1} \,\|\, 1-\pi_{1|0}\right) \le D\left(P_{X|H_1} \,\|\, P_{X|H_0}\right) \tag{12}$$
$$d\left(1-\pi_{1|0} \,\|\, \pi_{0|1}\right) \le D\left(P_{X|H_0} \,\|\, P_{X|H_1}\right) \tag{13}$$

where the binary divergence is denoted by

$$d(a \,\|\, b) = a \log \frac{a}{b} + (1-a) \log \frac{1-a}{1-b}. \tag{14}$$

Suppose now that a joint distribution P_{XY} is specified with marginals P_X and P_Y. Testing whether (X,Y) are dependent (according to the joint distribution P_{XY}) or are independent (with marginals P_X and P_Y) incurs in error probabilities (15) and (16) shown at the bottom of the page, that must satisfy

$$d\left(\pi_{\mathsf{i|d}} \,\|\, 1-\pi_{\mathsf{d|i}}\right) \le I(X;Y) \tag{17}$$
$$d\left(1-\pi_{\mathsf{d|i}} \,\|\, \pi_{\mathsf{i|d}}\right) \le L(X;Y). \tag{18}$$

Notice that the observation can be a vector of dimension n and the upper bounds in (17)–(18) become then $I(X^n;Y^n)$ and $L(X^n;Y^n)$. Fig. 2 shows the region (17)–(18) for $n=8$. As n grows, the region tends to a rectangle with corner point given by $(I(X;Y), L(X;Y))$.

2) Asymptotic Setting: The points in the region (17)–(18) that can actually be achieved can be conveniently characterized in the asymptotic regime of $n \to \infty$. Chernoff's result [18] (commonly referred to as Stein's lemma) states that if we observe n independent and identically distributed (i.i.d.) realizations of X and design the best hypothesis test such that $\pi_{0|1} \le \delta$, then the minimum $\pi_{1|0}$ satisfies

$$\lim_{n\to\infty} \frac{1}{n} \log \frac{1}{\pi_{1|0}} = D\left(P_{X|H_1} \,\|\, P_{X|H_0}\right) \tag{19}$$

where it is assumed that $D\left(P_{X|H_1} \,\|\, P_{X|H_0}\right) < \infty$.

If we now define hypothesis H_0 to denote a dependent joint distribution $(X,Y) \sim P_{XY}$ and hypothesis H_1 to denote an independent distribution $(X,Y) \sim P_X P_Y$, then for the best hypothesis test upon observing n i.i.d. realizations of (X,Y) such that $\pi_{\mathsf{d|i}} \le \delta$

$$\lim_{n\to\infty} \frac{1}{n} \log \frac{1}{\pi_{\mathsf{i|d}}} = L(X;Y) \tag{20}$$

provided that $L(X;Y) < \infty$. Analogously, for the best hypothesis test such that $\pi_{\mathsf{i|d}} \le \delta$, we have (cf. [19, Example 11.5.3])

$$\lim_{n\to\infty} \frac{1}{n} \log \frac{1}{\pi_{\mathsf{d|i}}} = I(X;Y). \tag{21}$$

These two achievable exponents of the probabilities correspond to the points $(I(X;Y), 0)$ and $(0, L(X;Y))$ in the plane of points $\left(\frac{1}{n}\log\frac{1}{\pi_{\mathsf{d|i}}}, \frac{1}{n}\log\frac{1}{\pi_{\mathsf{i|d}}}\right)$ as depicted in Fig. 2. The corner point $(I(X;Y), L(X;Y))$ can also be achieved in the sequential setting as outlined below.

The foregoing operational characterization serves to illustrate why the lautum information of a nondeterministic random variable with itself is infinite. Consider the special case where $X=Y$, i.e., the test between P_{XX} and $P_X \times P_X$. The test that declares P_{XX} if both components of all the observed pairs coincide, and $P_X \times P_X$ otherwise, achieves $\pi_{\mathsf{i|d}} = 0$ and $\pi_{\mathsf{d|i}} \to 0$. Note that according to (20) this requires that $L(X;X) = \infty$.

The corner point $(I(X;Y), L(X;Y))$ of the region in (17)–(18) is, in general, not achievable even asymptotically. It can be achieved in the alternative setup of sequential hypothesis testing introduced by Wald in his 1945 seminal paper [14], where the number of observations is allowed to depend on previous observations. In particular, [14] showed that the set of possible pairs $\left(\pi_{0|1}, \pi_{1|0}\right)$ must satisfy

$$d\left(\pi_{0|1} \,\|\, 1-\pi_{1|0}\right) \le \bar{n}_1 D\left(P_{X|H_1} \,\|\, P_{X|H_0}\right) \tag{22}$$
$$d\left(1-\pi_{1|0} \,\|\, \pi_{0|1}\right) \le \bar{n}_0 D\left(P_{X|H_0} \,\|\, P_{X|H_1}\right) \tag{23}$$

where $\bar{n}_i = \mathbb{E}\left[N \mid H_i\right]$ denotes the average number of required observations N under hypothesis i. In fact, (22)–(23) are achieved with equality in the sequential asymptotic setting of error probabilities going to zero as a direct consequence of Berk's result [20].

$$\pi_{\mathsf{d|i}} = \Pr\left[\text{decide } (X,Y) \text{ dependent} \mid (X,Y) \text{ are independent}\right] \tag{15}$$
$$\pi_{\mathsf{i|d}} = \Pr\left[\text{decide } (X,Y) \text{ independent} \mid (X,Y) \text{ are dependent}\right] \tag{16}$$

[4] Unless the logarithm basis is indicated, it can be chosen arbitrarily as long as both sides of the equation have the same units.

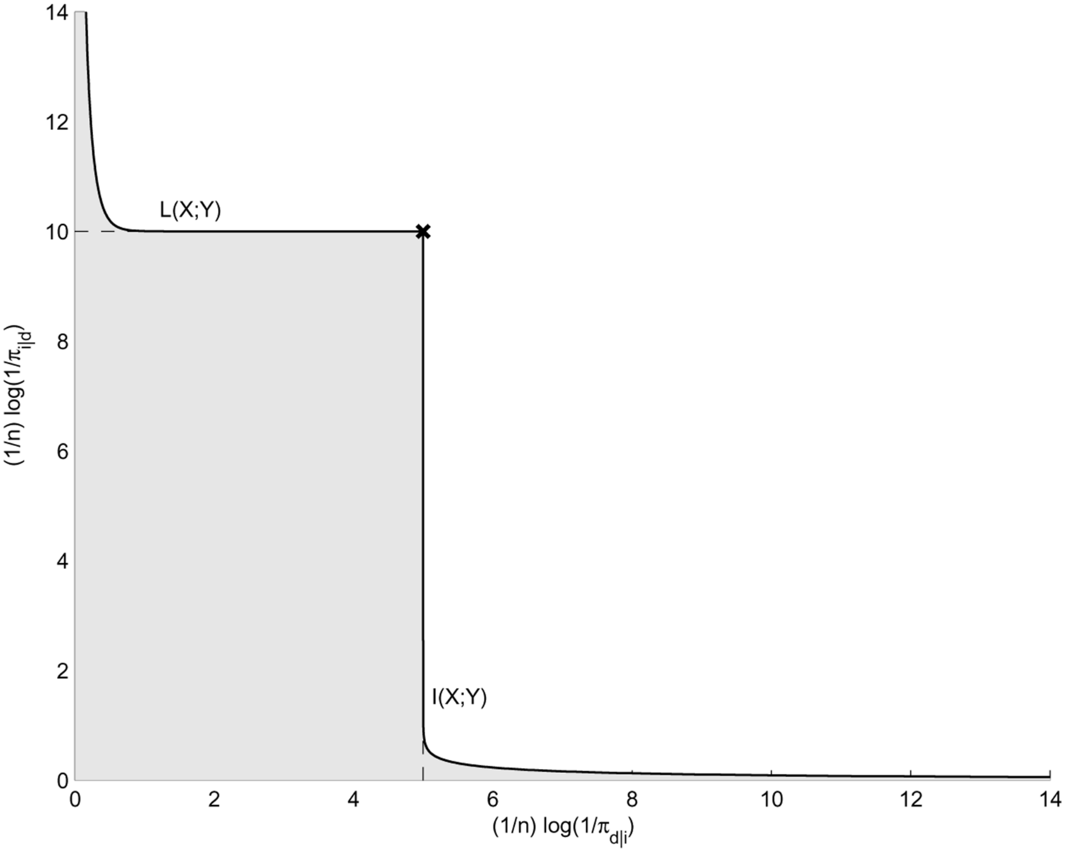

Fig. 2. Region where $\left(\pi_{\mathsf{d}|\mathsf{i}}, \pi_{\mathsf{i}|\mathsf{d}}\right)$ must lie when the observation has dimension $n = 8$ for $I(X;Y) = 5$ and $L(X;Y) = 10$.

Theorem 1: Assuming that both divergences are nonzero and finite, there exists a sequence of sequential hypothesis tests (indexed by k) that achieve

$$\lim_{k\to\infty} \frac{1}{N_0^{(k)}} \log \frac{1}{\pi_{0|1}^{(k)}} = D\left(P_{X|H_0} \,\|\, P_{X|H_1}\right) \quad P_{X|H_0} - \text{a.s.} \tag{24}$$

$$\lim_{k\to\infty} \frac{1}{N_1^{(k)}} \log \frac{1}{\pi_{1|0}^{(k)}} = D\left(P_{X|H_1} \,\|\, P_{X|H_0}\right) \quad P_{X|H_1} - \text{a.s.} \tag{25}$$

where the random stopping time $N_i^{(k)}$ denotes the number of required samples under hypothesis i. Equations (24)–(25) also hold replacing $N_i^{(k)}$ by its expectation $\bar{n}_i^{(k)}$.

Proof: See Appendix A. □

Accordingly, the best sequential hypothesis tests of dependence satisfy (a.s.)

$$\lim_{k\to\infty} \frac{1}{N_{\mathsf{i}}^{(k)}} \log \frac{1}{\pi_{\mathsf{i}|\mathsf{d}}^{(k)}} = L(X;Y) \tag{26}$$

$$\lim_{k\to\infty} \frac{1}{N_{\mathsf{d}}^{(k)}} \log \frac{1}{\pi_{\mathsf{d}|\mathsf{i}}^{(k)}} = I(X;Y). \tag{27}$$

Another large deviations result can be obtained from the method of types [21, eq. (II.6)]:

$$\Pr\left[(X^n, Y^n) \in \mathcal{T}_{P_X P_Y}^n\right] \approx \exp\left(-nL(X;Y)\right) \tag{28}$$

where the approximation is up to a polynomial factor (cf. [21, eq. (II.6)]) and $\mathcal{T}_{P_X P_Y}^n$ is the set of sequences with "product type" (finite input/output alphabets are assumed). Thus, lautum information determines the exponential decay of the probability that n realizations of a dependent pair of random variables will look independent.

B. Bayesian Testing of Independence

Consider the same setup as in the previous section where a random variable X is drawn from the distribution $P_{X|H_0}$ under hypothesis H_0 and from $P_{X|H_1}$ under hypothesis H_1. A Bayesian decision with minimum probability of error takes into account the priors of the hypotheses by comparing the following log-ratio of the posteriors of the n i.i.d. observations to a threshold:

$$l_n(x_1, \ldots, x_n) = \log \frac{\Pr[H_0]}{\Pr[H_1]} + \sum_{i=1}^{n} \log \frac{P_{X|H_0}(x_i)}{P_{X|H_1}(x_i)}. \tag{29}$$

Then, for i.i.d. $X_1, \ldots, X_n$ (cf. [22, Problem 3.6])

$$\frac{1}{n} l_n(X_1, \ldots X_n) \underset{\text{a.s.}}{\longrightarrow} \begin{cases} D\left(P_{X|H_0} \,\|\, P_{X|H_1}\right), & \text{if } H_0 \text{ is true} \\ -D\left(P_{X|H_1} \,\|\, P_{X|H_0}\right), & \text{if } H_1 \text{ is true} \end{cases} \tag{30}$$

provided that both divergences are finite.

In the independence testing Bayesian setup, where hypothesis H_0 denotes a dependent distribution $(X,Y) \sim P_{XY}$ and hypothesis H_1 denotes an independent distribution $(X,Y) \sim P_X P_Y$, it follows from the law of large numbers that

$$\begin{aligned} &\frac{1}{n} l_n\left((X_1,Y_1), \ldots (X_n,Y_n)\right) \\ &= \frac{1}{n} \log \frac{\Pr[H_0]}{\Pr[H_1]} + \frac{1}{n} \mathsf{i}\left((X_1,Y_1), \ldots, (X_n,Y_n)\right) \\ &\underset{\text{a.s.}}{\longrightarrow} \begin{cases} I(X;Y) & \text{if } (X,Y) \sim P_{XY} \\ -L(X;Y) & \text{if } (X,Y) \sim P_X P_Y. \end{cases} \end{aligned} \tag{31}$$

where $\mathsf{i}\left((x_1,y_1),\ldots,(x_n,y_n)\right)$ is the information density (e.g., [23]) of the observations $(x_1,y_1),\ldots,(x_n,y_n)$:

$$\mathsf{i}\left((x_1,y_1),\ldots,(x_n,y_n)\right)=\sum_{i=1}^{n}\log\frac{P_{XY}(x_i,y_i)}{P_X(x_i)P_Y(y_i)}. \tag{32}$$

In the simplest version of the random coding proof of the channel capacity theorem, the decoder compares the normalized information density achieved by each possible codeword to a threshold which is slightly below the mutual information; a message is declared if and only if it is the only one whose information density is above the threshold. According to (31), for the transmitted codeword the normalized information density converges to the mutual information, whereas for all the other codewords the normalized information density converges to minus the lautum information.

In the context of an authentication problem, where signatures are tested to decide whether they come from the same source, a more general independence hypothesis testing setup is analyzed in [9], where the signatures whose dependence/independence is tested are allowed to have memory. In that setting, the lautum information rate is defined as the natural counterpart to the mutual information rate, i.e., the limit of normalized lautum informations. The generalization of information density (32) to the setup with memory is shown in [9] to converge almost surely (normalized by n) to the corresponding information rates for a certain class of ergodic stationary processes.

C. Capacity Per Unit Cost of the Dependence-Test Channel

The capacity per unit cost [24] is defined similarly to the conventional capacity, except that the ratio of the logarithm of the number of codewords to their block length (rate) is replaced by the ratio of the logarithm of the number of codewords to their cost (rate per unit cost). The capacity per unit cost can be computed from the capacity–cost function $C(\beta)$, where β denotes the cost, by finding $\sup_{\beta>0} C(\beta)/\beta$ or, alternatively, as

$$\mathcal{C}=\sup_{P_X}\frac{I(X;Y)}{\mathbb{E}[b(X)]} \tag{33}$$

where $b(\cdot)$ is the cost function. In the important case where the input alphabet contains a zero-cost symbol (labeled as “0”) the capacity per unit cost is given by [24]

$$\mathcal{C}=\sup_{x}\frac{D\left(P_{Y|X=x}\parallel P_{Y|X=0}\right)}{b(x)} \tag{34}$$

where the supremum is over the input alphabet.

As an application of this result, consider now the *binary-input dependence-test channel* defined as a channel with binary input U with cost $b(u)=u$ and output (V,W) such that $(V,W)\sim P_{VW}$ for $U=0$ and $(V,W)\sim P_V P_W$ for $U=1$. The capacity per unit cost (34) is then equal to the lautum information:

$$\mathcal{C}=L(V;W). \tag{35}$$

Note that we can think of this setup as one in which it costs no “energy” to send dependent realizations of random variables while it takes some given expenditure to make them independent. An illustrative application is the case where V and W are the input/output of a BSC. At the expense of an increase in the noisiness of the channel, we can (covertly) communicate information to a third party who observes both V and W by switching off the link (and therefore making V and W independent) at certain times dependent on the covert message.

D. Description Length Penalty in Optimal Data Compression

Consider a source that generates i.i.d. symbols $X_1, X_2, \ldots$ drawn from the distribution P_X. The minimum expected codeword length to describe a symbol generated by the source with a binary prefix code is achieved by a Huffman code and is equal to $H(X)$ bits plus at most 1 bit (e.g., [19]). If the Huffman code is obtained for the distribution Q_X instead of P_X, there is a penalty ΔL in the expected codeword length approximately equal to $D(P_X \parallel Q_X)$ [19, Theorem 5.4.3].

In light of this result, it follows that if we have an optimum code designed for dependent symbols $(X,Y)\sim P_{XY}$, but the true source generates instead independent symbols $(X,Y)\sim P_X P_Y$, then the extra codeword length per symbol is

$$\Delta L = L(X;Y). \tag{36}$$

E. Doubling Rate Penalty in Kelly Gambling

Kelly’s operational characterization of mutual information [19] states that in the problem of gambling on the outcomes of a roulette (or horse racing) X_i, the increase in the doubling rate of wealth due to optimal use of side information Y_i is equal to $I(X;Y)$. If that optimum scheme is applied to the wrong roulette (whose outcomes are independent of the side information), then there is a penalty in doubling rate with respect to the gambling scheme that uses no side information. The penalty is equal to $L(X;Y)$. In other words, mutual information (resp., lautum information) quantifies the gain (resp., the loss) that accrues by assuming correctly (resp., incorrectly) that the side information is useful.

The previous observation for horse racing can be extended to the general stock market setup [19]. In that case, however, one can only obtain upper bounds in the increase/decrease of doubling rate.

III. Properties of Lautum Information

A. Basic Properties

Lautum information is indeed a *bone fide* measure of dependence. As an immediate consequence of its definition we have the following.

Theorem 2: (Nonnegativity of lautum information):

$$L(X;Y)=L(Y;X)\geq 0 \tag{37}$$

with equality if and only if X and Y are independent.

Proof: $D(P \parallel Q)\geq 0$ with equality if and only if $P=Q$. □

Nonnegativity also holds for the conditional version:

$$L(X;Y\mid Z)\triangleq D\left(P_{X|Z}P_{Y|Z}\parallel P_{XY|Z}\mid P_Z\right)\geq 0 \tag{38}$$

with equality if and only if the random variables form the Markov chain $X - Z - Y$.

Lautum information, however, does not satisfy the same chain rule as mutual information. For example, for the joint pmf on $\{0,1\}^3$

$$P_{X_1 X_2 Y}(x_1, x_2, y) = \begin{cases} \frac{1}{2} - 3\theta, & \text{if } x_1 = x_2 = y \\ \theta, & \text{otherwise} \end{cases} \tag{39}$$

where θ is a small positive parameter, it follows that (see Appendix B)

$$L(X_1, X_2; Y) > L(X_1; Y) + L(X_2; Y \mid X_1). \tag{40}$$

A sufficient condition for the chain rule

$$L(X_1, X_2; Y) = L(X_1; Y) + L(X_2; Y \mid X_1) \tag{41}$$

is that Y and X_1 be unconditionally independent. However, this condition is not necessary. For example, for a Markov chain, the chain rule is satisfied

$$\begin{aligned} L(X_i; X_{i-1} \cdots X_1) &= L(X_i; X_{i-1}) \\ &\quad + L(X_i; X_{i-2} \cdots X_1 \mid X_{i-1}) \qquad (42) \\ &= L(X_i; X_{i-1}). \qquad (43) \end{aligned}$$

If the Markov chain is stationary with transition probabilities $P_{kl} = \Pr[X_i = l \mid X_{i-1} = k]$, and stationary distribution denoted by μ_k, then

$$I(X_i; X_{i-1} \cdots X_1) = \sum_{k,l} \mu_k P_{kl} \log \frac{P_{kl}}{\mu_k} \tag{44}$$

$$L(X_i; X_{i-1} \cdots X_1) = \sum_{k,l} \mu_k \mu_l \log \frac{\mu_k}{P_{kl}}. \tag{45}$$

For a memoryless channel, the joint input–output mutual information is upper-bounded as (e.g., [19, Lemma 7.9.2])

$$I(X^n; Y^n) \le \sum_{i=1}^{n} I(X_i; Y_i). \tag{46}$$

Similarly, for a memoryless source, the mutual information is lower bounded as

$$I(X^n; Y^n) \ge \sum_{i=1}^{n} I(X_i; Y_i). \tag{47}$$

The lautum information between the inputs and outputs of a memoryless channel satisfies the counterpart of (47) (instead of (46)).

Theorem 3: (Lower bound on lautum information for a memoryless channel): If $P_{Y^n|X^n} = \prod_{i=1}^{n} P_{Y_i|X_i}$, then

$$L(X^n; Y^n) \ge \sum_{i=1}^{n} L(X_i; Y_i) \tag{48}$$

with equality if and only if $(Y_1, \ldots, Y_n)$ are independent.

Proof:

$$\begin{aligned} L(X^n; Y^n) &- \sum_{i=1}^{n} L(X_i; Y_i) \\ &= \int \log \frac{dP_{Y^n}}{dP_{Y^n|X^n}} dP_{X^n} dP_{Y^n} \\ &\quad - \sum_{i=1}^{n} \int \log \frac{dP_{Y_i}}{dP_{Y_i|X_i}} dP_{X_i} dP_{Y_i} \\ &= \int \left(\log \frac{dP_{Y^n}}{dP_{Y^n|X^n}} - \log \frac{dP_{Y_1}}{dP_{Y_1|X_1}} \cdot \cdots \cdot \frac{dP_{Y_n}}{dP_{Y_n|X_n}} \right) \\ &\quad \times dP_{X^n} dP_{Y^n} \\ &= \int \log \frac{dP_{Y^n}}{d(P_{Y_1} \times \cdots \times P_{Y_n})} dP_{Y^n} \\ &= D(P_{Y^n} \parallel P_{Y_1} \times \cdots \times P_{Y_n}) \\ &\ge 0. \end{aligned} \tag{49}$$

□

When the inputs (or outputs) are independent, we can find channels with memory for which (48) is satisfied with strict inequality and also channels for which (48) does not hold. For example, (48) is satisfied with strict inequality by choosing $P_{X^n Y^n}$ such that it vanishes at a single mass point (x^n, y^n) as this would make $L(X^n; Y^n) = +\infty$ whereas $\sum_{i=1}^{n} L(X_i; Y_i) < \infty$ since none of the marginals $P_{X_i Y_i}$ would vanish. An example for which (48) does not hold is given by the following binary-input binary-output channel under i.i.d. equiprobable inputs for $n > 1$ (see Appendix C):

$$P_{Y^n|X^n}(y^n \mid x^n) = \begin{cases} 1 - \delta, & \text{if } y^n = x^n \\ \delta / (2^n - 1), & \text{otherwise.} \end{cases} \tag{50}$$

The data processing inequality for a Markov chain $X - Y - Z$ states that $I(X;Y) \ge I(X;Z)$ and $I(Y;Z) \ge I(X;Z)$ (e.g., [19]). Interestingly, the same result holds for lautum information.

Theorem 4: (Data processing inequality): If $X - Y - Z$, then

$$L(X;Y) \ge L(X;Z) \tag{51}$$

with equality if and only if $X - Z - Y$.

Proof: We can expand the lautum information in two different ways:

$$\begin{aligned} L(X; Y, Z) &= L(X;Y) + D\left(P_{Z|Y} \parallel P_{Z|XY} \mid P_X P_Y\right) \qquad (52) \\ &= L(X;Z) + D\left(P_{Y|Z} \parallel P_{Y|XZ} \mid P_X P_Z\right). \qquad (53) \end{aligned}$$

Using $X - Y - Z$, we have $P_{Z|XY} = P_{Z|Y}$ which implies $D\left(P_{Z|Y} \parallel P_{Z|XY} \mid P_X P_Y\right) = 0$. Thus, (51) follows. Equality in (51) is satisfied if and only if $D\left(P_{Y|Z} \parallel P_{Y|XZ} \mid P_X P_Z\right) = 0$, i.e., if $X - Z - Y$. □

In fact, the data processing inequality of mutual and lautum informations can be obtained as particular cases of a more general version based on the f-divergence. The f-divergence is defined as [16]

$$D_f(P \parallel Q) = \int Q(\omega) f\left(\frac{P(\omega)}{Q(\omega)}\right) d\omega \tag{54}$$

where f is an arbitrary convex function. We can then define the f-information as

$$\begin{aligned} I_f(X;Y) &= D_f(P_X P_Y \parallel P_{XY}) \qquad (55) \\ &= \mathbb{E}\left[f\left(\frac{P_X(X) P_Y(Y)}{P_{XY}(X,Y)} \right) \right]. \qquad (56) \end{aligned}$$

Observe that mutual and lautum informations are particular cases of f-informations with convex functions defined on the positive real line $f_1(u) = -\log u$ and $f_2(u) = u \log u$, respectively: $I(X;Y) = I_{f_1}(X;Y)$ and $L(X;Y) = I_{f_2}(X;Y)$. Theorem 4 is a special case of the following result.

Theorem 5: [16] (Generalized data processing inequality): If $X - Y - Z$, then

$$I_f(X;Y) \geq I_f(X;Z) \tag{57}$$

with equality if and only if $X - Z - Y$.

A consequence of the data processing inequality in Theorem 4 is that lautum information, like mutual information, is impervious to deterministic one-to-one transformations. For example, $L(X;X+N)$ does not depend on the mean of the input.

Theorem 6: The lautum information $L(X;Y)$ is: i) a concave function of P_X for fixed $P_{Y|X}$, and ii) a convex function of $P_{Y|X}$ for fixed P_X.

Proof: The following shows the convexity in $P_{Y|X}$:

$$\begin{aligned} L(X;Y) &= D\left(P_Y \parallel P_{Y|X} \mid P_X\right) \\ &= D\left(\int P_{Y|X} dP_X \parallel P_{Y|X} \mid P_X\right) \end{aligned} \tag{58}$$

which is convex in $P_{Y|X}$ by convexity of divergence $D(P \parallel Q)$ in (P,Q).

The concavity in P_X can be shown as follows. Let

$$U = \begin{cases} 1, & \text{w.p. } \alpha \\ 0, & \text{w.p. } 1-\alpha, \end{cases} \tag{59}$$

X_0 and X_1 be two independent random variables, and note that $P_{X_U} = (1-\alpha)P_{X_0} + \alpha P_{X_1}$.

Consider the Markov chain $(U, X_0, X_1) - X_U - Y$. By the data processing theorem we have

$$\begin{aligned} L(X_U;Y) &\geq L(U, X_0, X_1; Y) \\ &= L(X_0, X_1; Y \mid U) + L(U;Y) \\ &\geq L(X_0, X_1; Y \mid U) \\ &= (1-\alpha) L(X_0, X_1; Y \mid U=0) \\ &\quad + \alpha L(X_0, X_1; Y \mid U=1) \\ &= (1-\alpha) L(X_0;Y) + \alpha L(X_1;Y). \end{aligned} \tag{60}$$

□

Theorem 7: Consider a discrete memoryless channel with additive noise, i.e.,

$$Y_i = X_i \oplus N_i \tag{61}$$

where N_i is an i.i.d. noise process with marginal distribution P_N and the addition $\oplus$ is that of a finite field defined on the finite input/output alphabet $\mathcal{A}$. Let P_U stand for the equiprobable distribution on $\mathcal{A}$. Then

$$\max_X I(X;Y) = D(P_N \parallel P_U) \tag{62}$$

$$\max_X L(X;Y) = D(P_U \parallel P_N) \tag{63}$$

$$\lim_{n\to\infty} \frac{1}{n} \max_{X^n} L(X^n;Y^n) = D(P_U \parallel P_N) + D(P_N \parallel P_U) \tag{64}$$

where the maximizing distribution in (62)–(63) is P_U and in (64) is a distribution with marginals equal to P_U such that $X_1 = \cdots = X_n$ with probability one.

Proof: The capacity result in (62) is well known and usually written as $\log|\mathcal{A}| - H(N)$.

To show (63), note that any two input distributions related by a cyclic shift achieve the same lautum information. Fix now a distribution that maximizes lautum information, and take a mixture with equal weights $(1/|\mathcal{A}|)$ of all its cyclic shifts. By the concavity result in Theorem 6, the mixture input distribution (which is equiprobable) cannot attain lower lautum information than the one we started with. Finally, since the output distribution corresponding to equiprobable input symbols is also equiprobable, (63) readily follows from the definition of lautum information.

To show (64), recall that in the proof of Theorem 3 we used the memorylessness of the channel to show that any n dimensional input distribution satisfies

$$L(X^n;Y^n) = \sum_{i=1}^n L(X_i;Y_i) + D\left(P_{Y^n} \parallel P_{Y_1} \times \cdots \times P_{Y_n}\right) \tag{65}$$

$$= \sum_{i=1}^n L(X_i;Y_i) + \sum_{i=1}^n H(Y_i) - H(Y^n) \tag{66}$$

$$\leq \sum_{i=1}^n L(X_i;Y_i) + \sum_{i=1}^n H(Y_i) - H(Y^n|X^n) \tag{67}$$

$$\leq nD(P_U \parallel P_N) + n\log|\mathcal{A}| - nH(N) \tag{68}$$

$$= nD(P_U \parallel P_N) + nD(P_N \parallel P_U). \tag{69}$$

On the other hand, let us consider the n–dimensional input distribution Q_{X^n} whose marginals are equiprobable and such that $X_1 = \cdots = X_n$ with probability one. This distribution achieves

$$L(X^n;Y^n) = nD(P_U \parallel P_N) + n\log|\mathcal{A}| - H(Y^n) \tag{70}$$

$$\geq nD(P_U \parallel P_N) + n\log|\mathcal{A}| - H(Y^n|X_1) - H(X_1) \tag{71}$$

$$= nD(P_U \parallel P_N) + n\log|\mathcal{A}| - nH(N) - H(X_1) \tag{72}$$

$$= nD(P_U \parallel P_N) + nD(P_N \parallel P_U) - \log|\mathcal{A}| \tag{73}$$

□

B. Variational Characterizations of Lautum Information

It is well known that mutual information satisfies the variational relations

$$\begin{aligned} I(X;Y) &= D(P_{XY} \parallel P_X P_Y) \\ &= \inf_{Q_Y} D(P_{XY} \parallel P_X Q_Y) \end{aligned} \tag{74}$$

$$= \inf_{Q_X Q_Y} D(P_{XY} \parallel Q_X Q_Y). \tag{75}$$

The counterpart for lautum information would obtain $D(P_X P_Y \parallel P_{XY})$ as the infimum of $D(P_X Q_Y \parallel P_{XY})$ over

Q_Y. However, this is not true as shown by the following counterexample: choose P_{XY} such that it vanishes only at a single mass point (x_0, y_0) (this implies positive marginals) and Q_Y such that $Q_Y(y_0) = 0$; this leads to $D(P_X P_Y \parallel P_{XY}) = \infty$ while $D(P_X Q_Y \parallel P_{XY}) < \infty$. Regarding a possible counterpart to (75) note that $\inf_{Q_{XY}} D(P_X P_Y \parallel Q_{XY}) = 0$.

Lautum information satisfies the following variational characterization:

$$L(X;Y) = D(P_X P_Y \parallel P_{XY}) = \inf_{Q_X} D\left(P_X P_Y \parallel Q_X P_{Y|X}\right) \tag{76}$$

where $Q_X P_{Y|X}$ stands for the joint distribution $Q_X(x) P_{Y|X}(y \mid x)$. Equation (76) follows from

$$D\left(P_X P_Y \parallel Q_X P_{Y|X}\right) = D\left(P_Y \parallel P_{Y|X} \mid P_X\right) + D(P_X \parallel Q_X). \tag{77}$$

Other useful identities in the context of lautum information are

$$\begin{aligned} D(Q_X Q_Y \parallel P_{XY}) &= D\left(Q_X \parallel P_{X|Y} \mid Q_Y\right) + D(Q_Y \parallel P_Y) \\ &= D\left(Q_Y \parallel P_{Y|X} \mid Q_X\right) + D(Q_X \parallel P_X) \end{aligned} \tag{78}$$

and

$$\begin{aligned} D\left(Q_Y \parallel P_{Y|X} \mid Q_X\right) = D\left(Q_X \parallel P_{Y|X} \mid Q_Y\right) \\ + D(Q_Y \parallel P_Y) + H(Q_X) - H(P_Y). \end{aligned} \tag{79}$$

C. Bounds on Information Measures

Since both mutual information and lautum information are defined as divergences, they inherit the properties and bounds known for divergence (e.g., [17], [19], [25]–[27], and references therein). In particular the Csiszár–Pinsker–Kemperman inequality [25, p. 58], [19, Lemma 11.6.1], $D(P \parallel Q) \geq \frac{\log e}{2} d^2(P;Q)$, where $d(P;Q) \triangleq \|P - Q\|_1$, implies

$$\frac{\log e}{2} V^2(X;Y) \leq \min\{I(X;Y), L(X;Y)\} \tag{80}$$

where $V(X;Y)$ is the *variational distance* between the distributions P_{XY} and $P_X P_Y$, defined as the l_1-norm:

$$V(X;Y) \triangleq \|P_{XY} - P_X P_Y\|_1. \tag{81}$$

An account of alternative lower bounds on divergence is given in [27].

An inequality relating mutual information and expectation was given in [28, Lemma 4.4] (with an improvement in a scaling factor in [29]). The following result extends the inequality also to lautum information with a much simpler proof.

Theorem 8: Let X, Y be random variables with X taking discrete values in the unit interval $[-1, 1]$. Then[5]

$$\begin{aligned} \frac{\log e}{2} \mathbb{E}^2\left[\left|\mathbb{E}[X \mid Y] - \mathbb{E}[X \mid f(Y)]\right|\right] \\ \leq \min\{I(X;Y \mid f(Y)), L(X;Y \mid f(Y))\} \end{aligned} \tag{82}$$

where $f(\cdot)$ is a deterministic function.

Proof: From the Csiszár–Pinsker–Kemperman inequality and $d(P;Q) \geq \left|\sum_x xP(x) - xQ(x)\right|$, we have that

$$D(P \parallel Q) \geq \frac{\log e}{2} \left|\sum_x xP(x) - xQ(x)\right|^2. \tag{83}$$

Particularizing (83) to $P = P_{X|Y'=y'}$ and $Q = P_{X|Y=y,Y'=y'}$, where Y' is, for now, an arbitrary random variable, we get

$$\begin{aligned} &D\left(P_{X|Y'=y'} \parallel P_{X|Y=y,Y'=y'}\right) \\ &\geq \frac{\log e}{2} \left|\mathbb{E}[X \mid Y'=y'] - \mathbb{E}[X \mid Y=y, Y'=y']\right|^2. \end{aligned} \tag{84}$$

Taking now expectation with respect to $P_{YY'} = P_{Y'} P_{Y|Y'}$, we obtain

$$\begin{aligned} &D\left(P_{X|Y'} P_{Y|Y'} \parallel P_{XY|Y'} \mid P_{Y'}\right) \\ &\geq \frac{\log e}{2} \mathbb{E}\left[\left|\mathbb{E}[X \mid Y'] - \mathbb{E}[X \mid Y, Y']\right|^2\right] \qquad (85) \\ &\geq \frac{\log e}{2} \mathbb{E}^2\left[\left|\mathbb{E}[X \mid Y'] - \mathbb{E}[X \mid Y, Y']\right|\right] \qquad (86) \end{aligned}$$

where the second inequality follows from Jensen's inequality. Defining $Y' = f(Y)$ leads to (82). The proof for mutual information follows identical steps. □

Regarding the comparison between both measures of information, it turns out that lautum information is larger than or equal to mutual information for many cases of interest such as the input/output of the BSC and the Gaussian channel. In general, however, this is not true as shown by the following counterexample with joint distribution given by

P_{XY}	$Y=0$	$Y=1$
$X=0$	0.96	0.01
$X=1$	0.01	0.02

(87)

which corresponds to a lautum information smaller than the mutual information: $L(X;Y) = 0.0584$ and $I(X;Y) = 0.0865$.

Both mutual information and lautum information are measures of the dependence between random variables. However, in cases where the distributions are unknown and their information measures are estimated through a universal estimator [30], lautum information may provide a more useful gauge of dependence than mutual information. For example, if any of the random variables has a small entropy, mutual information will also be small and may be indistinguishable from the estimation noise whereas lautum information need not be small (as it is not upper-bounded by the entropy). The following example illustrates the different sensitivities of lautum/mutual information to parameters in the joint distribution.

Example 1: Consider a binary-symmetric Markov chain with transition probability $p = 0.0002$ through a BSC with crossover probability δ. The following table quantifies the values of lautum and mutual informations (computed with the method in [31] and references therein):

δ	$L(X;Y)$	$I(X;Y)$
0.01	3.2412	0.0027
0.05	1.9062	0.0026
0.10	1.2637	0.0025

(88)

[5] In [28], [29], Y is additionally assumed to be discrete.

D. Lower Bounds on Error Probability

Fano's inequality lower-bounds the mutual information between random variables that take values on the same finite set with cardinality M as follows (e.g., [32]):

$$I(X;Y) \geq \Pr[X=Y]\log M - h(\Pr[X=Y]) \tag{89}$$

which holds as long as either X or Y is equiprobable, where $h(x)$ is the binary entropy function. A stronger lower bound on the mutual information between random variables that take values on the same set is [32, Theorem 3]

$$I(X;Y) \geq d\left(\Pr[X=Y] \,\|\, \Pr\left[\bar{X}=\bar{Y}\right]\right) \tag{90}$$

where the binary divergence function is defined in (14).

A similar lower bound for the lautum information is given next.

Theorem 9: If X and Y take values on the same set, then

$$L(X;Y) \geq d\left(\Pr\left[\bar{X}=\bar{Y}\right] \,\|\, \Pr[X=Y]\right) \tag{91}$$

where $\bar{X}$ and $\bar{Y}$ are independent and have the same marginal distributions as X and Y, respectively.

Proof: Application of the data processing theorem for divergence ("processing reduces divergence") to a processor that takes as input (x,y) and outputs $1\{x=y\}$ under the different input distributions $P_X P_Y$ and P_{XY}. □

In the special case in which either X or Y is equiprobable on a finite set of cardinality M, Theorem 9 becomes

$$\begin{aligned} L(X;Y) &\geq d\left(\frac{1}{M} \,\|\, \Pr[X=Y]\right) \\ &= \frac{1}{M}\log\frac{1}{\Pr[X=Y]} \\ &\quad + \left(1-\frac{1}{M}\right)\log\frac{1}{\Pr[X\neq Y]} - h\left(\frac{1}{M}\right) \\ &\geq \left(1-\frac{1}{M}\right)\log\frac{1}{\Pr[X\neq Y]} - h\left(\frac{1}{M}\right). \end{aligned} \tag{92}$$

The bound in (92) leads to the following upper bound on the reliability function (evaluated for the BSC at the end of Section IV).

Theorem 10: Consider the transmission of a code with block length n and rate R over a channel $\left\{P_{Y^n|X^n}\right\}_{n=1}^{\infty}$. Let $P_e(n,R)$ be the minimum error probability for any such code and denote the channel reliability function (cf. [33], [34]) by

$$E(R) \triangleq \liminf_{n\to\infty}\frac{1}{n}\log\frac{1}{P_e(n,R)}. \tag{93}$$

Then

$$E(R) \leq E\left(0^+\right) \leq \liminf_{n\to\infty}\frac{1}{n}\sup_{P_{X^n}} L(X^n;Y^n). \tag{94}$$

Proof: From the following Markov chain denoting the communication process over the channel

$$X - X^n(W) - Y^n - \hat{W}(Y^n),$$

letting $M=2^{nR}$, and (92) we have

$$\begin{aligned} \liminf_{n\to\infty}\frac{1}{n}L\left(W;\hat{W}\right) &\geq \liminf_{n\to\infty}\frac{1}{n}\left(1-\frac{1}{2^{nR}}\right)\log\frac{1}{\Pr\left[W\neq\hat{W}\right]} \\ &\quad - \lim_{n\to\infty}\frac{1}{n}h\left(\frac{1}{2^{nR}}\right) \end{aligned} \tag{95}$$

$$= \liminf_{n\to\infty}\frac{1}{n}\log\frac{1}{\Pr\left[W\neq\hat{W}\right]}. \tag{96}$$

Since this is true for any code, it also holds for the best code with minimum error probability $\Pr\left[W\neq\hat{W}\right] = P_e(n,R)$:

$$\liminf_{n\to\infty}\frac{1}{n}L\left(W;\hat{W}\right) \geq \liminf_{n\to\infty}\frac{1}{n}\log\frac{1}{P_e(n,R)} = E(R). \tag{97}$$

Using now the data processing inequality for lautum information (Theorem 4) we have

$$\frac{1}{n}L\left(W;\hat{W}\right) \leq \frac{1}{n}L(X^n;Y^n) \leq \frac{1}{n}\sup_{P_{X^n}} L(X^n;Y^n) \tag{98}$$

and the result follows. □

The upper bound in (94) need not be tight as is illustrated in Section IV for the BSC.

IV. Lautum Information for the BSC

Theorem 11: The input/output lautum information for the BSC with crossover probability δ and $P_X(1)=p$ is

$$\begin{aligned} L(X;Y) &= d(q\,\|\,\delta) + h(q) - H(Y) \\ &= p(1-2\delta) \\ &\quad \times \left(2(1-p)\log\frac{1-\delta}{\delta} + \log\frac{p(1-2\delta)+\delta}{1-p(1-2\delta)-\delta}\right) \\ &\quad - d(\delta\,\|\,p(1-2\delta)+\delta) \end{aligned} \tag{99}$$

where $q=\delta+2p(1-p)(1-2\delta)$.

Proof: First note that $P_Y(1)=p(1-2\delta)+\delta$ and

$$\begin{aligned} q &= \Pr\left[\bar{X}\neq\bar{Y}\right] \\ &= pP_Y(0) + (1-p)P_Y(1) \\ &= \delta + 2p(1-p)(1-2\delta). \end{aligned}$$

Then

$$\begin{aligned} L(X;Y) &= (1-p)P_Y(0)\log\frac{1}{1-\delta} + (1-p)P_Y(1)\log\frac{1}{\delta} \\ &\quad + pP_Y(0)\log\frac{1}{\delta} + pP_Y(1)\log\frac{1}{1-\delta} - H(Y) \\ &= q\log\frac{1}{\delta} + (1-q)\log\frac{1}{1-\delta} - H(Y) \\ &= d(q\,\|\,\delta) + h(q) - H(Y) \end{aligned} \tag{100}$$

and (99) follows from

$$\begin{aligned} H(Y) &= d(\delta\,\|\,p(1-2\delta)+\delta) + h(\delta) \\ &\quad - p(1-2\delta)\log\frac{p(1-2\delta)+\delta}{1-p(1-2\delta)-\delta}. \end{aligned} \tag{101}$$

□

Theorem 12: For the BSC, $L(X;Y) \geq I(X;Y)$.

Proof: The proof is equivalent to the recent refinement of Pinsker's inequality in [35, Theorem 2.1] particularized to the binary case (see Appendix D). □

Particularizing Theorem 7 we obtain that equiprobable inputs maximize the lautum information for the BSC.[6]

Theorem 13: The maximal lautum information for the BSC satisfies

$$\max_{X} L(X;Y) = \frac{1}{2}\log\frac{1}{4\delta(1-\delta)} \tag{102}$$

$$\lim_{n\to\infty}\max_{X^n}\frac{1}{n}L(X^n;Y^n) = \left(\frac{1}{2}-\delta\right)\log\frac{1-\delta}{\delta}. \tag{103}$$

The expression on the right-hand side of (102) has appeared in a number of problems in the literature, e.g., as the error exponent at channel capacity of a particular transmission system over the BSC with feedback [36] and as the exponent in an upper bound on the error probability for linear block codes over the BSC [37, Sec. 2.9].

The bound in Theorem 10 is loose for the BSC. The channel reliability function for the BSC with crossover probability δ is [33, Sec. 5.8], [34, Problem 10.13]

$$E(0) = \frac{1}{4}\log\frac{1}{4\delta(1-\delta)}, \tag{104}$$

whereas the right-hand side of (94) is given by (103) which is strictly larger except for $\delta = 0.5$.

V. The Gaussian Case

A. Lautum Information for the Gaussian Channel

Consider a general discrete-time linear vector Gaussian channel represented by the following vector signal model with n_T transmit dimensions and n_R receive dimensions:

$$\boldsymbol{Y} = \mathbf{H}\boldsymbol{X} + \boldsymbol{N} \tag{105}$$

where all quantities are complex-valued, $\boldsymbol{X}$ is the n_T-dimensional transmitted vector arbitrarily distributed (not necessarily Gaussian), $\mathbf{H}$ is the $n_R \times n_T$ matrix that denotes the linear transformation undergone by the signal, $\boldsymbol{Y}$ is the n_R-dimensional received vector, and $\boldsymbol{N}$ is an n_R-dimensional proper complex Gaussian noise vector independent of $\boldsymbol{X}$. The input and the noise covariance matrices are

$$\boldsymbol{\Sigma} = \mathbb{E}\left[(\boldsymbol{X}-\mathbb{E}[\boldsymbol{X}])(\boldsymbol{X}-\mathbb{E}[\boldsymbol{X}])^\dagger\right]$$

and

$$\boldsymbol{\Phi} = \mathbb{E}\left[(\boldsymbol{N}-\mathbb{E}[\boldsymbol{N}])(\boldsymbol{N}-\mathbb{E}[\boldsymbol{N}])^\dagger\right]$$

respectively.

Theorem 14: Consider the Gaussian signal model in (105) where $\boldsymbol{X}$ is arbitrarily distributed with zero mean.[7] Then, the mutual information and lautum information are given by[8]

$$I(\boldsymbol{X};\boldsymbol{Y}) = \mathrm{Tr}\left(\boldsymbol{\Phi}^{-1}\mathbf{H}\boldsymbol{\Sigma}\mathbf{H}^\dagger\right)\log e - D(P_{\boldsymbol{Y}} \,\|\, P_{\boldsymbol{N}}) \tag{106}$$

$$L(\boldsymbol{X};\boldsymbol{Y}) = \mathrm{Tr}\left(\boldsymbol{\Phi}^{-1}\mathbf{H}\boldsymbol{\Sigma}\mathbf{H}^\dagger\right)\log e + D(P_{\boldsymbol{Y}} \,\|\, P_{\boldsymbol{N}}). \tag{107}$$

Proof: For mutual information, the proof is straightforward

$$\begin{aligned} &D\left(P_{\boldsymbol{Y}|\boldsymbol{X}} \,\|\, P_{\boldsymbol{Y}} \mid P_{\boldsymbol{X}}\right) \\ &\qquad = D\left(P_{\boldsymbol{Y}|\boldsymbol{X}} \,\|\, P_{\boldsymbol{N}} \mid P_{\boldsymbol{X}}\right) - D(P_{\boldsymbol{Y}} \,\|\, P_{\boldsymbol{N}}) \end{aligned} \tag{108}$$

and the application of the divergence between two proper complex Gaussian distributions [38]

$$\begin{aligned} &D(\mathcal{CN}(\mathbf{m}_1,\boldsymbol{\Sigma}_1), \mathcal{CN}(\mathbf{m}_0,\boldsymbol{\Sigma}_0)) \\ &\quad = \log\det(\boldsymbol{\Sigma}_0) - \log\det(\boldsymbol{\Sigma}_1) \\ &\qquad + (\mathbf{m}_1-\mathbf{m}_0)^\dagger\boldsymbol{\Sigma}_0^{-1}(\mathbf{m}_1-\mathbf{m}_0)\log e \\ &\qquad + \mathrm{Tr}\left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\Sigma}_1 - \mathbf{I}\right)\log e. \end{aligned} \tag{109}$$

For lautum information, the proof is slightly more involved and one needs to resort to the fact that both $P_{\boldsymbol{N}}$ and $P_{\boldsymbol{Y}|\boldsymbol{X}}$ are Gaussian distributed with the same covariance matrix $\boldsymbol{\Phi}$ and means $\mathbf{0}$ and $\mathbf{H}\boldsymbol{X}$ (the results obtained hold verbatim for an arbitrary noise mean), respectively:

$$\begin{aligned} &D\left(P_{\boldsymbol{Y}} \,\|\, P_{\boldsymbol{Y}|\boldsymbol{X}} \mid P_{\boldsymbol{X}}\right) \\ &\quad = \mathbb{E}\left[\log\frac{P_{\boldsymbol{N}}(\bar{\boldsymbol{Y}})}{P_{\boldsymbol{Y}|\boldsymbol{X}}(\bar{\boldsymbol{Y}} \mid \bar{\boldsymbol{X}})}\right] + D(P_{\boldsymbol{Y}} \,\|\, P_{\boldsymbol{N}}) &(110)\\ &\quad = \mathbb{E}\left[-\bar{\boldsymbol{Y}}^\dagger\Phi^{-1}\bar{\boldsymbol{Y}} + \left(\bar{\boldsymbol{Y}}-\mathbf{H}\bar{\boldsymbol{X}}\right)^\dagger\Phi^{-1}\left(\bar{\boldsymbol{Y}}-\mathbf{H}\bar{\boldsymbol{X}}\right)\right] \\ &\qquad + D(P_{\boldsymbol{Y}} \,\|\, P_{\boldsymbol{N}}) &(111)\\ &\quad = \mathbb{E}\left[\bar{\boldsymbol{X}}^\dagger\mathbf{H}^\dagger\Phi^{-1}\mathbf{H}\bar{\boldsymbol{X}} - \bar{\boldsymbol{X}}^\dagger\mathbf{H}^\dagger\Phi^{-1}\bar{\boldsymbol{Y}} - \bar{\boldsymbol{Y}}^\dagger\Phi^{-1}\mathbf{H}\bar{\boldsymbol{X}}\right] \\ &\qquad + D(P_{\boldsymbol{Y}} \,\|\, P_{\boldsymbol{N}}) &(112)\\ &\quad = \mathrm{Tr}\left(\mathbf{H}^\dagger\Phi^{-1}\mathbf{H}\boldsymbol{\Sigma}\right)\log e + D(P_{\boldsymbol{Y}} \,\|\, P_{\boldsymbol{N}}). &(113) \end{aligned}$$

□

Theorem 15: For the Gaussian channel, $L(\boldsymbol{X};\boldsymbol{Y}) \geq I(\boldsymbol{X};\boldsymbol{Y})$.

Proof: According to Theorem 14, the difference is a divergence $L(\boldsymbol{X};\boldsymbol{Y}) - I(\boldsymbol{X};\boldsymbol{Y}) = 2D(P_{\boldsymbol{Y}} \,\|\, P_{\boldsymbol{N}})$. □

For Gaussian noise, it is well known that mutual information is maximized, under a second-order moment constraint, when the input is Gaussian. In the case of lautum information, the opposite happens in light of the relation

$$L(\boldsymbol{X};\boldsymbol{Y}) + I(\boldsymbol{X};\boldsymbol{Y}) = 2\mathrm{Tr}\left(\boldsymbol{\Phi}^{-1}\mathbf{H}\boldsymbol{\Sigma}\mathbf{H}^\dagger\right)\log e \tag{114}$$

which, according to Theorem 14, holds for any $\boldsymbol{X}$ (not necessarily Gaussian) with covariance $\boldsymbol{\Sigma}$.

[6] The particularization of (99) to equiprobable inputs appears in [9].

[7] If $\mathbb{E}[\boldsymbol{X}] \neq \mathbf{0}$, the term $D(P_{\boldsymbol{Y}} \,\|\, P_{\boldsymbol{N}})$ appearing in (106) and (107) must be replaced by $D(P_{\tilde{\mathbf{Y}}} \,\|\, P_{\boldsymbol{N}}) = D(P_{\boldsymbol{Y}} \,\|\, P_{\boldsymbol{N}}) - \left\|\Phi^{-1/2}\mathbf{H}\mathbb{E}[\boldsymbol{X}]\right\|^2$, where $\tilde{\boldsymbol{Y}} = \boldsymbol{Y} - \mathbf{H}\mathbb{E}[\boldsymbol{X}]$ which has the same distribution as $\boldsymbol{Y}$ but with the mean of $\boldsymbol{X}$ removed.

[8] For the case of real-valued random variables, (106)–(107) require a factor $1/2$ in front of the term with the trace.

An interesting property of mutual information is the saddle-point characterization of the Gaussian distribution[9]

$$I\left(\boldsymbol{X};\mathbf{H}\boldsymbol{X}+\boldsymbol{N}_G\right) \leq I\left(\boldsymbol{X}_G;\mathbf{H}\boldsymbol{X}_G+\boldsymbol{N}_G\right) \leq I\left(\boldsymbol{X}_G;\mathbf{H}\boldsymbol{X}_G+\boldsymbol{N}\right) \qquad (115)$$

where $\boldsymbol{X}_G$ and $\boldsymbol{N}_G$ are Gaussian and independent, and $\boldsymbol{X}$ and $\boldsymbol{N}$ follow arbitrary distributions with the same second-order moments as the Gaussian counterparts. Lautum information does not admit a similar saddle-point characterization. As previously argued, Gaussian inputs minimize lautum information:

$$L\left(\boldsymbol{X}_G;\mathbf{H}\boldsymbol{X}_G+\boldsymbol{N}_G\right) \leq L\left(\boldsymbol{X};\mathbf{H}\boldsymbol{X}+\boldsymbol{N}_G\right). \qquad (116)$$

However, the other required inequality for the saddle-point characterization is not satisfied; simply by choosing $P_{\boldsymbol{N}}$ vanishing at some set of nonzero measure we obtain $L\left(\boldsymbol{X}_G;\mathbf{H}\boldsymbol{X}_G+\boldsymbol{N}\right) = +\infty$. For some other examples, however, the inequality is satisfied such as with a Laplacian noise (with sufficiently small noise power).

Theorem 16: Consider the Gaussian signal model in (105) where $\boldsymbol{X}$ is Gaussian distributed. Then, the mutual information and lautum information are given by

$$I\left(\boldsymbol{X};\boldsymbol{Y}\right) = \log\det\left(\mathbf{I}+\mathbf{\Phi}^{-1}\mathbf{H}\mathbf{\Sigma}\mathbf{H}^{\dagger}\right) \qquad (117)$$

$$L\left(\boldsymbol{X};\boldsymbol{Y}\right) = 2\mathrm{Tr}\left(\mathbf{\Phi}^{-1}\mathbf{H}\mathbf{\Sigma}\mathbf{H}^{\dagger}\right)\log e - \log\det\left(\mathbf{I}+\mathbf{\Phi}^{-1}\mathbf{H}\mathbf{\Sigma}\mathbf{H}^{\dagger}\right). \qquad (118)$$

Proof: Particularize Theorem 14 using from (109)

$$D\left(P_{\boldsymbol{Y}} \parallel P_{\boldsymbol{N}}\right) = -\log\det\left(\mathbf{I}+\mathbf{\Phi}^{-1}\mathbf{H}\mathbf{\Sigma}\mathbf{H}^{\dagger}\right) + \mathrm{Tr}\left(\mathbf{\Phi}^{-1}\mathbf{H}\mathbf{\Sigma}\mathbf{H}^{\dagger}\right)\log e. \qquad (119)$$

□

B. Lautum Information for Jointly Gaussian Random Variables

This subsection evaluates the mutual information and the lautum information between two proper complex vector Gaussian random variables: $\boldsymbol{X} \sim \mathcal{CN}\left(\mathbf{m}_x,\mathbf{\Sigma}_x\right)$ and $\boldsymbol{Y} \sim \mathcal{CN}\left(\mathbf{m}_y,\mathbf{\Sigma}_y\right)$.

The joint and the product distributions $P_{\boldsymbol{XY}}$ and $P_{\boldsymbol{X}}P_{\boldsymbol{Y}}$ can be respectively encompassed in the random variables $\boldsymbol{Z}$ and $\bar{\boldsymbol{Z}}$ obtained by stacking the vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ such that $\boldsymbol{Z} \sim \mathcal{CN}\left(\mathbf{m}_z,\mathbf{\Sigma}_z\right)$ and $\bar{\boldsymbol{Z}} \sim \mathcal{CN}\left(\mathbf{m}_z,\bar{\mathbf{\Sigma}}_z\right)$ where the covariance matrices are given by

$$\mathbf{\Sigma}_z = \begin{bmatrix} \mathbf{\Sigma}_x & \mathbf{\Sigma}_{xy} \\ \mathbf{\Sigma}_{yx} & \mathbf{\Sigma}_y \end{bmatrix} \quad \text{and} \quad \bar{\mathbf{\Sigma}}_z = \begin{bmatrix} \mathbf{\Sigma}_x & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_y \end{bmatrix}. \qquad (120)$$

Assuming nonsingularity of the covariance matrices, the mutual information can be easily evaluated as [40]

$$I\left(\boldsymbol{X};\boldsymbol{Y}\right) = -\log\det\left(\mathbf{I}-\mathbf{\Sigma}_y^{-1}\mathbf{\Sigma}_{yx}\mathbf{\Sigma}_x^{-1}\mathbf{\Sigma}_{xy}\right). \qquad (121)$$

The lautum information can be similarly evaluated.

Theorem 17: Let $\left(\boldsymbol{X},\boldsymbol{Y}\right)$ be two vector joint Gaussian random variables with covariance matrix

$$\begin{bmatrix} \mathbf{\Sigma}_x & \mathbf{\Sigma}_{xy} \\ \mathbf{\Sigma}_{yx} & \mathbf{\Sigma}_y \end{bmatrix}$$

(assuming $\mathbf{\Sigma}_x > \mathbf{0}$ and $\mathbf{\Sigma}_y > \mathbf{0}$). Then

$$L\left(\boldsymbol{X};\boldsymbol{Y}\right) = \log\det\left(\mathbf{I}-\mathbf{\Sigma}_y^{-1}\mathbf{\Sigma}_{yx}\mathbf{\Sigma}_x^{-1}\mathbf{\Sigma}_{xy}\right) + 2\mathrm{Tr}\left(\left(\mathbf{I}-\mathbf{\Sigma}_y^{-1}\mathbf{\Sigma}_{yx}\mathbf{\Sigma}_x^{-1}\mathbf{\Sigma}_{xy}\right)^{-1}-\mathbf{I}\right)\log e. \qquad (122)$$

Proof: The proof follows from the application of the divergence between two proper complex Gaussian distributions (109) and some additional algebraic manipulations. Alternatively, the proof follows easily from Theorem 16 for the signal model $\boldsymbol{Y} = \mathbf{H}\boldsymbol{X}+\boldsymbol{N}$ by properly choosing the channel as $\mathbf{H} = \mathbf{\Sigma}_{yx}\mathbf{\Sigma}_x^{-1}$ and the noise covariance matrix as $\mathbf{\Sigma}_n = \mathbf{\Sigma}_y - \mathbf{\Sigma}_{yx}\mathbf{\Sigma}_x^{-1}\mathbf{\Sigma}_{xy}$. □

Particularizing Theorem 17 to Toeplitz matrices and using standard asymptotic results [41], it follows that

$$L\left(X;Y\right) = \int\left(\log\left(1-\left|\rho\left(f\right)\right|^2\right) + 2\left(\frac{1}{1-\left|\rho\left(f\right)\right|^2}-1\right)\log e\right)df \qquad (123)$$

where $\rho\left(f\right)$ is the frequency-dependent normalized covariance given by

$$\rho\left(f\right) = \frac{\mathbb{E}\left[\left(X\left(f\right)-m_x\left(f\right)\right)\left(Y\left(f\right)-m_y\left(f\right)\right)^*\right]}{\sigma_x\left(f\right)\sigma_y\left(f\right)}. \qquad (124)$$

Another particular case of Theorem 17 was given in [9] for the signal model

$$\begin{bmatrix} X \\ Y \end{bmatrix} = Z\begin{bmatrix} 1 \\ 1 \end{bmatrix} + \begin{bmatrix} W_1 \\ W_2 \end{bmatrix}. \qquad (125)$$

Theorem 18: If $\left(\boldsymbol{X};\boldsymbol{Y}\right)$ are jointly Gaussian random vectors, then $L\left(\boldsymbol{X};\boldsymbol{Y}\right) \geq I\left(\boldsymbol{X};\boldsymbol{Y}\right)$.

Proof: It follows from Theorem 14, by noting that any jointly Gaussian vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ admit the formula $\boldsymbol{Y} = \mathbf{H}\boldsymbol{X}+\boldsymbol{N}$ with $\boldsymbol{N}$ independent of $\boldsymbol{X}$. □

APPENDIX

A. Proof of Theorem 1

Achievability is shown with the sequential probability ratio test (SPRT) [14]. Denoting the upper and lower thresholds a_1 and $-a_0$, respectively, it follows that the result can be shown by letting $a^{(k)} = \min\left(a_0^{(k)},a_1^{(k)}\right)$ be an increasing sequence going to infinity. From [20, Theorem 2.1],

$$\lim\frac{N_0^{(k)}}{a_0^{(k)}} = \lim\frac{\bar{n}_0^{(k)}}{a_0^{(k)}} = \frac{1}{D\left(P_{X|H_0} \parallel P_{X|H_1}\right)} \qquad (126)$$

$$\lim\frac{N_1^{(k)}}{a_1^{(k)}} = \lim\frac{\bar{n}_1^{(k)}}{a_1^{(k)}} = \frac{1}{D\left(P_{X|H_1} \parallel P_{X|H_2}\right)}, \qquad (127)$$

[9] Mutual information also admits a saddle-point characterization of the exponential distribution [39]; lautum information, again, does not share such a characterization (observe that with an additive exponential noise, $L\left(\boldsymbol{X};\boldsymbol{Y}\right) = \infty$ for any nondeterministic input).

and, from [20, Theorem 2.2]

$$\lim \frac{1}{a_0} \log \frac{1}{\pi_{0|1}^{(k)}} = \lim \frac{1}{a_1} \log \frac{1}{\pi_{1|0}^{(k)}} = 1. \tag{128}$$

Therefore

$$\begin{aligned}\lim \frac{1}{N_0^{(k)}} \log \frac{1}{\pi_{0|1}^{(k)}} &= \lim \frac{a_0^{(k)}}{N_0^{(k)}} \frac{1}{a_0^{(k)}} \log \frac{1}{\pi_{0|1}^{(k)}} &(129)\\ &= \lim \frac{a_0^{(k)}}{N_0^{(k)}} \lim \frac{1}{a_0^{(k)}} \log \frac{1}{\pi_{0|1}^{(k)}} &(130)\\ &= D\left(P_{X|H_0} \parallel P_{X|H_1}\right) &(131)\end{aligned}$$

and similarly for (25).

B. Counterexample of the Chain Rule for Lautum Information

From the joint distribution defined in (39), the following marginals and conditionals easily follow: $P_{X_1}(0) = P_{X_1}(1) = 1/2$, $P_Y(0) = P_Y(1) = 1/2$,

$$P_{X_1X_2}(x_1, x_2) = \begin{cases} 0.5 - 2\theta, & \text{if } x_1 = x_2 \\ 2\theta, & \text{otherwise,} \end{cases} \tag{132}$$

$$P_{Y|X_1}(y \mid x_1) = P_{X_2|X_1}(y \mid x_1) = \begin{cases} 1 - 4\theta, & \text{if } y = x_1 \\ 4\theta, & \text{otherwise,} \end{cases} \tag{133}$$

and

$$P_{Y|X_1X_2}(y \mid x_1, x_2) = \begin{cases} (0.5 - 3\theta)/(0.5 - 2\theta), & \text{if } y = x_1 = x_2 \\ \theta/(0.5 - 2\theta), & \text{if } y \neq x_1 = x_2 \\ 0.5, & \text{otherwise.} \end{cases} \tag{134}$$

Then, the following approximation can be obtained for sufficiently small θ:

$$\begin{aligned}&L(X_1, X_2; Y) - (L(X_1; Y) + L(X_2; Y \mid X_1)) \\ &\quad = \frac{1}{2}\log 2 - (\log 4e)\,\theta + 2\theta(1 - 8\theta)\log\theta + O\left(\theta^2\right) &(135)\\ &\quad \to \frac{1}{2}\log 2. &(136)\end{aligned}$$

C. Example of $L(X^n; Y^n) < \sum_{i=1}^n L(X_i; Y_i)$ With Independent Inputs

From the problem setup, it follows that $P_{X^n}(x^n) = 2^{-n} = P_{Y^n}(y^n)$, $P_{X_i}(x_i) = 1/2 = P_{Y_i}(y_i)$, and

$$P_{Y_i|X_i}(y_i \mid x_i) = \begin{cases} (1 - \delta) + \delta\left(2^{n-1} - 1\right)/(2^n - 1) & \text{if } y_i = x_i \\ \delta 2^{n-1}/(2^n - 1) & \text{otherwise.} \end{cases} \tag{137}$$

It is then straightforward to evaluate

$$L(X^n; Y^n) = \frac{1}{2^n}\left[\log \frac{1}{(1 - \delta)2^n} + (2^n - 1)\log \frac{2^n - 1}{\delta 2^n}\right] \tag{138}$$

and

$$L(X_i; Y_i) = \frac{1}{2}\left[\log \frac{2^n - 1}{\delta 2^n} + \log \frac{1/2}{(1 - \delta) + \delta(2^{n-1} - 1)/(2^n - 1)}\right]. \tag{139}$$

The following inequality can then be verified for $n > 1$:

$$nL(X_i; Y_i) - L(X^n; Y^n) > 0. \tag{140}$$

In fact, it can be checked that the left-hand side of (140) grows linearly with n, as $n \to \infty$.

D. Proof of Theorem 12

From Theorem 11 and using $I(X;Y) = H(Y) - h(\delta)$ and

$$d(q \parallel \delta) + h(q) = h(\delta) + 2p(1 - p)(1 - 2\delta)\log \frac{1 - \delta}{\delta},$$

we have after some algebra

$$\begin{aligned}&L(X;Y) - I(X;Y) \\ &\quad = -2d(\delta \parallel p(1 - 2\delta) + \delta) + 2(1 - 2\delta) \\ &\qquad \times p\left((1 - p)\log \frac{1 - \delta}{\delta} + \log \frac{1 - p(1 - 2\delta) - \delta}{p(1 - 2\delta) + \delta}\right) &(141)\end{aligned}$$

which is positive from a direct application of Lemma 2 below with $a = \delta$ and $b = p(1 - 2\delta) + \delta$.

Lemma 1: [35, Theorem 2.1] (Lower bound on the binary divergence)

$$d(a \parallel b) \geq \frac{(a - b)^2}{1 - 2b}\log \frac{1 - b}{b}. \tag{142}$$

Lemma 2: (Upper bound on the binary divergence)

$$(a - b)\left(\log \frac{1 - b}{b} - \frac{1 - b - a}{1 - 2a}\log \frac{1 - a}{a}\right) \geq d(a \parallel b). \tag{143}$$

Proof: First, note the identity

$$d(a \parallel b) + d(b \parallel a) = (a - b)\left(\log \frac{1 - b}{b} - \log \frac{1 - a}{a}\right). \tag{144}$$

Then, from Lemma 1

$$d(a \parallel b) + d(b \parallel a) \geq \frac{(a - b)^2}{1 - 2b}\log \frac{1 - b}{b} + d(b \parallel a) \tag{145}$$

or equivalently

$$(a - b)\left(\log \frac{1 - b}{b} - \log \frac{1 - a}{a}\right) \geq \frac{(a - b)^2}{1 - 2b}\log \frac{1 - b}{b} + d(b \parallel a) \tag{146}$$

which reduces to (143) after rearranging terms and interchanging a and b. □

REFERENCES

[1] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423 and 623, Oct. 1948.

[2] L. Paninski, "Estimation of entropy and mutual information," *Neural Comput.*, vol. 15, no. 6, pp. 1191–1253, Jun. 2003.

[3] Z. Dawy, B. Goebel, J. Hagenauer, C. Andreoli, T. Meitinger, and J. Mueller, "Gene mapping and marker clustering using Shannon's mutual information," *IEEE Trans. Computat. Biol. Bioinform.*, vol. 3, no. 1, pp. 47–56, Jan.-Mar. 2006.

[4] K. Torkkola, "Feature extraction by nonparametric mutual information maximization," *J. Mach. Learning Res.*, vol. 3, pp. 1415–1438, Mar. 2003.

[5] N. Slonim, G. S. Atwal, G. Tkacik, and W. Bialek, "Information based clustering," *Proc. Nat. Acad. Sci. (USA)*, vol. 102, pp. 18297–18302, 2005.

[6] J. P. Pluim and J. B. Maintz, "Mutual-information-based registration of medical images: a survey," *IEEE Trans. Med. Imag.*, vol. 22, no. 8, pp. 986–1004, Aug. 2003.

[7] A. Kaltchenko, "Algorithms for estimating information distance with applications to bioinformatics and linguistics," in *Proc. IEEE Canadian Conf. Electrical and Computer Engineering (CCECE 2004)*, Niagara Falls, ON, Canada, May 2004.

[8] M. Zaffalon and M. Hutter, "Robust feature selection by mutual information distributions," in *Proc. 18th Annu. Conf. Uncertainty in Artificial Intelligence (UAI-02)*, San Francisco, CA, 2002, pp. 577–584.

[9] J. A. O'Sullivan and N. A. Schmid, "Performance analysis of physical signature authentication," *IEEE Trans. Inf. Theory*, vol. 47, no. 7, pp. 3034–3039, Nov. 2001.

[10] A. Hero, B. Ma, O. Michel, and J. Gorman, "Applications of entropic spanning graphs," *IEEE Signal Process. Mag.*, vol. 19, no. 5, pp. 85–95, Sep. 2002.

[11] R. W. Yeung, "A new outlook on Shannon's information measures," *IEEE Trans. Inf. Theory*, vol. 37, no. 3, pp. 466–474, May 1991.

[12] R. W. Yeung, *A First Course in Information Theory*. Norwell, MA: Kluwer Academic/Plenum, 2002.

[13] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, pp. 79–86, 1951.

[14] A. Wald, "Sequential tests of statistical hypotheses," *Ann. Math. Statist.*, vol. 16, no. 2, pp. 117–186, Jun. 1945.

[15] H. Jeffreys, "An invariant form for the prior probability in estimation problems," *Proc. Roy. Soc. Lon., Ser. A*, vol. 186, pp. 453–461, 1946.

[16] I. Csiszár, "A class of measures of informativity of observation channels," *Periodica Mathematica Hungarica*, vol. 2, no. 1–4, pp. 191–213, 1972.

[17] S. Kullback, *Information Theory and Statistics*. New York: Dover, 1968.

[18] H. Chernoff, "Large-sample theory: Parametric case," *Ann. Math. Statist.*, vol. 27, no. 1, pp. 1–22, Mar. 1956.

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

[20] R. H. Berk, "Some asymptotic aspects of sequential analysis," *Ann. Statist.*, vol. 1, no. 6, pp. 1126–1138, Nov. 1973.

[21] I. Csiszár, "The method of types," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2505–2523, Oct. 1998.

[22] S. Verdú, *Multiuser Detection*. New York: Cambridge Univ. Press, 1998.

[23] S. Verdú and T. S. Han, "A general formula for channel capacity," *IEEE Trans. Inf. Theory*, vol. 40, no. 4, pp. 1147–1157, Jul. 1994.

[24] S. Verdú, "On channel capacity per unit cost," *IEEE Trans. Inf. Theory*, vol. 36, no. 5, pp. 1019–1030, Sep. 1990.

[25] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic, 1981.

[26] F. Topsoe, "Some inequalities for information divergence and related measures of discrimination," *IEEE Trans. Inf. Theory*, vol. 46, no. 4, pp. 1602–1609, Jul. 2000.

[27] A. A. Fedotov, P. Harremöes, and F. Topsoe, "Refinements of Pinsker's inequality," *IEEE Trans. Inf. Theory*, vol. 49, no. 6, pp. 1491–1498, Jun. 2003.

[28] T. Tao, "Szeméredi's regularity lemma revisited," *Contrib. Discr. Math.*, vol. 1, no. 1, pp. 8–28, 2006.

[29] R. Ahlswede, "The final form of Tao's inequality relating conditional expectation and conditional mutual information," *Adv. Math. of Commun.*, vol. 1, no. 2, pp. 239–242, 2007.

[30] S. Verdú, "Universal estimation of information measures," in *Proc. 2005 IEEE Information Theory Workshop*, Rotorua, New Zealand, Aug. 2005.

[31] S. Yang and A. Kavčić, "Capacity of partial response channels," in *Handbook on Coding and Signal Processing for Recording Systems*. Boca Raton, FL: CRC, 2004, ch. 13.

[32] T. S. Han and S. Verdú, "Generalizing the Fano inequality," *IEEE Trans. Inf. Theory*, vol. 40, no. 4, pp. 1247–1251, Jul. 1994.

[33] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.

[34] R. E. Blahut, *Principles and Practice of Information Theory*. Reading, MA: Addison-Wesley, 1987.

[35] E. Ordentlich and M. J. Weinberger, "A distribution dependent refinement of Pinsker's inequality," *IEEE Trans. Inf. Theory*, vol. 51, no. 5, pp. 1836–1840, May 2005.

[36] M. Horstein, "Sequential transmission using noiless feedback," *IEEE Trans. Inf. Theory*, vol. IT–9, no. 3, pp. 136–143, Jul. 1963.

[37] A. J. Viterbi and J. K. Omura, *Principles of Digital Communication and Coding*. New York: McGraw-Hill, 1979.

[38] S. Verdú, "Spectral efficiency in the wideband regime," *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1319–1343, Jun. 2002.

[39] S. Verdú, "The exponential distribution in information theory," *Probl. Inf. Transm.*, vol. 32, no. 1, pp. 86–95, 1996.

[40] M. S. Pinsker, *Information and Information Stability of Random Variables and Processes*. San Francisco, CA: Holden-Day, 1964.

[41] R. M. Gray, "On the asymptotic eigenvalue distribution of Toeplitz matrices," *IEEE Trans. Inf. Theory*, vol. IT-18, no. 6, pp. 725–730, Nov. 1972.