

# Lautum Information

Daniel P. Palomar and Sergio Verdú

Dept. of Electrical Engineering  
Princeton University  
Princeton, NJ 08544, USA  
Email: {danielp,verdu}@princeton.edu

**Abstract**—A popular way to measure the degree of dependence between two random variables is with mutual information, defined as the divergence between the joint and product-of-marginal distributions. We introduce an alternative measure of dependence we refer to as *lautum information*: the divergence between the product-of-marginal and joint distributions. Some operational characterizations and properties are provided for this alternative measure of information.

## I. INTRODUCTION

A popular measure of the degree of dependence between  $X$  and  $Y$  is the *mutual information* defined as the divergence between the joint and product-of-marginal distributions:

$$I(X;Y) \triangleq D(P_{XY} \parallel P_X P_Y) \quad (1)$$

$$= D(P_{Y|X} \parallel P_Y | P_X). \quad (2)$$

In this paper we explore an alternative measure of dependence where the roles of the joint and product-of-marginal distributions are swapped. We define the *lautum information*<sup>1</sup> between  $X$  and  $Y$  as

$$L(X;Y) \triangleq D(P_X P_Y \parallel P_{XY}) \quad (3)$$

$$= D(P_Y \parallel P_{Y|X} | P_X). \quad (4)$$

Mutual information is, arguably, the most important specialization of divergence. Even before Kullback and Leibler introduced  $D(P||Q)$  [1], Jeffreys [2] introduced the symmetrized form  $D(P||Q) + D(Q||P)$ . Yet (3) appears to have remained unexplored. This paper provides several operational characterizations for the lautum information and derives a number of useful properties.

*Notation:* We define  $(\bar{X}, \bar{Y})$  to be independent random variables with the same marginals as the random variables  $(X, Y)$ . Unless the logarithm basis is indicated, it can be chosen arbitrarily as long as both sides of the equation have the same units.

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.

<sup>1</sup>The word *lautum* (“elegant” in latin) is the reverse spelling of *mutual*.

## II. OPERATIONAL CHARACTERIZATIONS OF LAUTUM INFORMATION

### A. Non-Bayesian Testing of Independence

Suppose we observe  $n$  independent identically distributed realizations of pairs of random variables  $(X_i, Y_i)$  where  $X_i$  and  $Y_i$  have known marginal distributions. We need to decide whether the pairs are drawn from a given joint distribution or they are independent. Using Stein’s lemma [3], for the best hypothesis test upon observing  $n$  iid realizations of  $(X, Y)$  such that  $\Pr[\text{decide } (X, Y) \sim P_{XY} | (X, Y) \sim P_X P_Y \text{ is true}] \leq \delta$ , the complementary error probability is characterized by the lautum information as

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log 1/\Pr[\text{decide } (X, Y) \sim P_X P_Y | (X, Y) \sim P_{XY}] = L(X;Y) \quad (5)$$

provided that  $L(X;Y) < \infty$ . Observe that swapping the hypothesis we obtain that for the best hypothesis test such that  $\Pr[\text{decide } (X, Y) \sim P_X P_Y | (X, Y) \sim P_{XY} \text{ is true}] \leq \delta$ , the complementary error probability is characterized by the mutual information as

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log 1/\Pr[\text{decide } (X, Y) \sim P_{XY} | (X, Y) \sim P_X P_Y] = I(X;Y). \quad (6)$$

A result similar to (5) can be obtained from the method of types [4, Lem. II.2]:

$$\Pr[(X^n, Y^n) \in T_{P_X P_Y}^n] \simeq \exp(-nL(X;Y)) \quad (7)$$

where  $T_{P_X P_Y}^n$  is the set of sequences with “product type” (finite input/output alphabets are assumed). Thus, lautum information controls the exponential decay of the probability that a dependent pair of random variables will behave as being independent.

### B. Bayesian Testing of Independence

In the independence testing Bayesian setup, where a prior for the probability that  $X$  and  $Y$  are independent is defined, the minimum probability of error test compares the following log-likelihood ratio of the posterior of the  $n$  iid observations

to a threshold

$$l_n((x_1, y_1), \dots, (x_n, y_n)) = \log \frac{\Pr[X, Y \text{ dependent}]}{\Pr[X, Y \text{ independent}]} + \sum_{i=1}^n \log \frac{P_{XY}(x_i, y_i)}{P_X(x_i)P_Y(y_i)}. \quad (8)$$

Then, (c.f. [5, Prob. 3.6])

$$\frac{1}{n} l_n \xrightarrow{\text{a.s.}} \begin{cases} I(X; Y) & \text{if } (X, Y) \sim P_{XY} \\ -L(X; Y) & \text{if } (X, Y) \sim P_X P_Y. \end{cases} \quad (9)$$

#### C. Capacity per Unit Cost of the Dependence-Test Channel

The capacity per unit cost [6] is defined similarly to the conventional capacity, except that the ratio of the logarithm of the number of codewords to their blocklength (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  $\beta$  denotes the cost, by finding  $\sup_{\beta > 0} C(\beta)/\beta$  or, alternatively, as

$$C = \sup_{P_X} \frac{I(X; Y)}{\mathbb{E}[b(X)]} \quad (10)$$

where  $b(\cdot)$  is the cost function. Interestingly, as shown in [6], in the important case where the input alphabet contains a zero-cost symbol (labeled as “0”) the capacity per unit cost is given by

$$C = \sup_x \frac{D(P_{Y|X=x} \| P_{Y|X=0})}{b(x)} \quad (11)$$

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  $U$  and output  $(X, Y)$  such that  $(X, Y) \sim P_{XY}$  for  $U = 0$  and  $(X, Y) \sim P_X P_Y$  for  $U = 1$ . Defining the cost as  $b(u) = u$ , then the capacity per unit cost (11) is equal to the lautum information:

$$C = L(X; Y). \quad (12)$$

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.

The capacity of this channel can be upper bounded by

$$\frac{1}{4} (I(X; Y) + L(X; Y)). \quad (13)$$

#### D. Description Length Penalty in Optimal Data Compression

Consider a source  $P_X$ . It is well known that the minimum expected codeword length to describe a symbol generated by the source with a binary prefix code is equal to  $H(X)$  bits plus at most 1 bit. If the code is optimized for the distribution  $Q_X$  instead of  $P_X$ , there is a penalty in the expected codeword length  $\Delta L$  quantified by  $D(P_X \| Q_X)$  [7, Thm. 5.4.3].

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 expected codeword length per symbol  $\Delta L$  is given by the lautum information:

$$\Delta L = L(X; Y). \quad (14)$$

#### E. Doubling Rate Penalty in Optimal Portfolio Investment

Consider a stock market represented by the vector  $\mathbf{X} = (X_1, \dots, X_m)$ , where  $m$  is the number of stocks and  $X_i$  is the relative price of the  $i$ th stock (i.e., ratio of the price at the end of the day to the price at the beginning of the day). A portfolio  $\mathbf{b} = (b_1, \dots, b_m)$ , where  $b_i \geq 0$  and  $\sum_i b_i = 1$ , is an allocation of wealth across the stocks (i.e.,  $b_i$  is the fraction of one’s wealth invested in stock  $i$ ). The doubling rate of a stock market portfolio  $\mathbf{b}$  is defined as  $W(\mathbf{b}, \mathbf{X}) \triangleq \mathbb{E}[\log \langle \mathbf{b}, \mathbf{X} \rangle]$  and a portfolio that achieves the maximum of  $W(\mathbf{b}, \mathbf{X})$  is called a log-optimal portfolio (c.f. [7]). The justification for the definition of the doubling rate is that, if  $X_1, \dots, X_n$  are iid drawn from  $P_X$ , then the wealth after  $n$  days using portfolio  $\mathbf{b}$  satisfies [7]

$$\frac{1}{n} \log \prod_{i=1}^n \langle \mathbf{b}, \mathbf{X}_i \rangle \rightarrow W \text{ a.s.} \quad (15)$$

Interestingly, it turns out that if  $X_1, \dots, X_n$  are iid drawn from  $P_X$  and the log-optimal portfolio is incorrectly designed for the distribution  $Q_X$ , then there is a penalty in the achieved doubling rate  $\Delta W$  upper-bounded by  $D(P_X \| Q_X)$  [7, Thm. 15.4.1].

Consider a market with two independent stocks  $(X, Y) \sim P_X P_Y$ . If a log-optimal portfolio is designed under the incorrect assumption that the stocks are dependent  $(X, Y) \sim P_{XY}$  (with the same marginals as the true distribution), then there is a loss in the doubling rate  $\Delta W$  upper-bounded by the lautum information:

$$\Delta W \leq L(X; Y). \quad (16)$$

Consider now a market with  $m$  stocks  $\mathbf{X} = (X_1, \dots, X_m)$  drawn from a distribution  $P_X$  independent of the side information given by the random variable  $Y$ . If a log-optimal portfolio is designed under the incorrect assumption that the stocks are dependent on  $Y$  according to  $P_{XY}$ , then the loss in the doubling rate  $\Delta W$  is upper-bounded by the lautum information:

$$\Delta W \leq L(\mathbf{X}; Y). \quad (17)$$

### 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

$$L(X; Y) = L(Y; X) \geq 0 \quad (18)$$

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

For a memoryless channel, it follows that the mutual information is upper bounded as (e.g., [7, Lem. 8.9.2])

$$I(X^n; Y^n) \leq \sum_{i=1}^n I(X_i; Y_i). \quad (19)$$

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

$$I(X^n; Y^n) \geq \sum_{i=1}^n I(X_i; Y_i). \quad (20)$$

The lautum information between the inputs and outputs of a memoryless channel satisfies the counterpart of (20) (instead of (19)):

*Theorem 1:* (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

$$L(X^n; Y^n) \geq \sum_{i=1}^n L(X_i; Y_i) \quad (21)$$

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

When the inputs (or outputs) are independent, we can find channels with memory for which (21) is satisfied with strict inequality and we can also find channels for which (21) does not hold.

The data processing inequality for a Markov chain  $X - Y - Z$  states that  $I(X; Y) \geq I(X; Z)$  and  $I(Y; Z) \geq I(X; Z)$  (e.g., [7]). Interestingly, the same result holds when the  $-\log \omega$  in the definition of mutual information is substituted by an arbitrary convex nonincreasing functional (under some technical conditions) such as  $e^{-\omega}$  [8]. Although the lautum information does not fall within that class of information measures (the equivalent functional would be  $\omega \log \omega$  which is convex but not monotonic), the data processing inequality holds as well.

*Theorem 2:* (Data processing inequality): If  $X - Y - Z$ , then

$$L(X; Y) \geq L(X; Z) \quad (22)$$

$$L(Y; Z) \geq L(X; Z). \quad (23)$$

A consequence of the data processing inequality in Theorem 2 is that lautum information, like mutual information, is insensitive to deterministic one-to-one transformations. For example,  $L(X; X + N)$  is insensitive to the mean of the input.

In parallel to mutual information we have (adapting the  $I(P, V)$  notation from [9]):

*Theorem 3:* Let

$$L(P_X, P_{Y|X}) = L(X; Y).$$

Then  $L(P_X, P_{Y|X})$  is concave in  $P_X$  and convex in  $P_{Y|X}$ .

Lautum information satisfies the following variational characterization:

$$L(X; Y) = D(P_X P_Y \parallel P_{XY}) = \inf_{Q_X} D(P_X P_Y \parallel Q_X P_{Y|X}) \quad (24)$$

where  $Q_X P_{Y|X}$  stands for the joint distribution  $Q_X(x) P_{Y|X}(y|x)$ . Another useful identity in the context

of lautum information is

$$\begin{aligned} D(Q_X \times Q_Y \parallel P_{XY}) &= D(Q_X \parallel P_{X|Y} | Q_Y) + D(Q_Y \parallel P_Y) \\ &= D(Q_Y \parallel P_{Y|X} | Q_X) + D(Q_X \parallel P_X). \end{aligned} \quad (25)$$

### B. 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., [10], [9], [7]). In particular, by the Csiszár-Pinsker-Kemperman inequality [9, p. 58]

$$\min \{I(X; Y), L(X; Y)\} \geq \frac{\log e}{2} V^2(X; Y) \quad (26)$$

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

$$V(X; Y) \triangleq \|P_{XY} - P_X P_Y\|_1. \quad (27)$$

For discrete input and output alphabets, the mutual information is upper bounded by the log of cardinalities of the input/output alphabets. Only when both input and output alphabets are continuous,  $I(X; Y)$  can be unbounded. The lautum information, however, can be unbounded even with discrete input/output alphabets: for example,  $L(X; X) = +\infty$  unless  $X$  is deterministic. As another example,  $L(X; Y) = +\infty$  if  $X$  and  $Y$  are connected through a discrete memoryless channel (e.g. a binary erasure channel) such that  $P_{Y|X}(y_0|x_0) = 0$  while  $P_X(x_0) > 0$  and  $P_Y(y_0) > 0$  for some  $(x_0, y_0)$ .

Regarding the comparison between both measures of information, it turns out that lautum information is larger than or equal to mutual information for some cases of interests such as the input/output of the binary symmetric channel (BSC) and the Gaussian channel. In a general case, however, this is not true as shown by the following counterexample with joint distribution given by  $P_{XY}(0, 0) = 0.96$ ,  $P_{XY}(1, 1) = 0.02$ ,  $P_{XY}(0, 1) = P_{XY}(1, 0) = 0.01$ . In this case, (in bits)

$$L(X; Y) = 0.0584, \quad I(X; Y) = 0.0865. \quad (28)$$

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 [11], 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).

### C. Lower Bounds on Error Probability

In this subsection, we give counterparts to Fano's inequality and its generalizations [12].

*Theorem 4:* If  $X$  and  $Y$  take values on the same set, then

$$L(X; Y) \geq d(\Pr[\bar{X} = \bar{Y}] \parallel \Pr[X = Y]) \quad (29)$$

where  $\bar{X}$  and  $\bar{Y}$  are independent and have the same marginal distributions as  $X$  and  $Y$ , respectively, and  $d(x \| y)$  is the binary divergence function defined as the continuous extension on  $[0, 1]^2$  of

$$d(x \| y) \triangleq x \log \frac{x}{y} + (1-x) \log \frac{1-x}{1-y}. \quad (30)$$

By using the lower bound on the binary divergence  $d(x \| y) \geq x \log \frac{1}{y} - h(x)$ , we can write

$$L(X; Y) \geq \Pr[\bar{X} = \bar{Y}] \log \frac{1}{\Pr[X = Y]} - h(\Pr[\bar{X} = \bar{Y}]) \quad (31)$$

where  $h(x)$  is the binary entropy function.

A special case is when either  $X$  or  $Y$  is equiprobable on a finite set of cardinality  $M$ . In this case,  $\Pr[\bar{X} = \bar{Y}] = 1/M$  and then it follows from Theorem 4 that

$$\begin{aligned} L(X; Y) &\geq \frac{1}{M} \log \frac{1}{M \Pr[X = Y]} + \left(1 - \frac{1}{M}\right) \log \frac{1 - 1/M}{\Pr[X \neq Y]} \\ &\geq \left(1 - \frac{1}{M}\right) \log \frac{1}{\Pr[X \neq Y]} - h(1/M). \end{aligned} \quad (32)$$

Observe that for large  $M$ , we can write the approximation

$$L(X; Y) \gtrsim \log \frac{1}{\Pr[X \neq Y]}. \quad (33)$$

One application of this result is the following. Consider an encoder/channel/decoder system, where a message  $W$  is encoded as  $X^n$ , received through the channel as  $Y^n$ , and decoded as  $\hat{W}$ . Then

$$\frac{1}{n} \log \frac{1}{\Pr[W \neq \hat{W}]} \lesssim \frac{1}{n} L(W; \hat{W}) \quad (34)$$

$$\leq \frac{1}{n} L(X^n; Y^n) \quad (35)$$

$$\leq \frac{1}{n} \sup_{P_{X^n}} L(X^n; Y^n) \quad (36)$$

where the second upper bound follows from the data processing inequality for lautum information (Theorem 2).

Defining  $P_e(n, R)$  as the minimum error probability for a code with blocklength  $n$  and rate  $R$ , we can upper bound the channel reliability function (c.f. [13], [14]) as

$$E(R) \triangleq \liminf_{n \rightarrow \infty} \frac{1}{n} \log \frac{1}{P_e(n, R)} \leq \liminf_{n \rightarrow \infty} \frac{1}{n} \sup_{P_{X^n}} L(X^n; Y^n) \quad (37)$$

However, this upper bound is loose as shown by the next example.

*Example 1:* The channel reliability function for the BSC with crossover probability  $\delta$  is [14, Prob. 10.13]

$$E(0) = \frac{1}{4} \log \frac{1}{4\delta(1-\delta)}, \quad (38)$$

whereas

$$\sup_{P_X} L(X; Y) = \frac{1}{2} \log \frac{1}{4\delta(1-\delta)} \quad (39)$$

and

$$\sum_{i=1}^n \sup_{P_{X_i}} L(X_i; Y_i) \leq \sup_{P_{X^n}} L(X^n; Y^n). \quad (40)$$

Therefore, the upper bound  $\liminf_{n \rightarrow \infty} \frac{1}{n} \sup_{P_{X^n}} L(X^n; Y^n)$  is off by at least a factor of two.

#### IV. 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:

$$\mathbf{Y} = \mathbf{H}\mathbf{X} + \mathbf{N} \quad (41)$$

where all quantities are complex-valued,  $\mathbf{X}$  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,  $\mathbf{Y}$  is the  $n_R$ -dimensional received vector, and  $\mathbf{N}$  is an  $n_R$ -dimensional proper complex Gaussian noise vector independent of  $\mathbf{x}$ . The input and the noise covariance matrices are  $\Sigma$  and  $\Phi$ , respectively.

*Theorem 5:* Consider the Gaussian signal model in (41) where  $\mathbf{X}$  is arbitrarily distributed with zero mean.<sup>2</sup> Then,<sup>3</sup>

$$I(\mathbf{X}; \mathbf{Y}) = \text{Tr}(\Phi^{-1} \mathbf{H} \Sigma \mathbf{H}^\dagger) \log e - D(P_Y \| P_N) \quad (42)$$

$$L(\mathbf{X}; \mathbf{Y}) = \text{Tr}(\Phi^{-1} \mathbf{H} \Sigma \mathbf{H}^\dagger) \log e + D(P_Y \| P_N). \quad (43)$$

An immediate corollary of Theorem 5 is that for the Gaussian channel lautum information is greater or equal than mutual information. Interestingly, the same relationship also holds for the BSC.<sup>4</sup>

It is well known that mutual information is maximized, given the first- and second-order moments and under Gaussian noise, when the input is Gaussian. In the case of lautum information, the opposite happens in light of the relation

$$L(\mathbf{X}; \mathbf{Y}) + I(\mathbf{X}; \mathbf{Y}) = 2 \text{Tr}(\Phi^{-1} \mathbf{H} \Sigma \mathbf{H}^\dagger) \log e. \quad (44)$$

It is remarkable that as measures of dependence between random variables, mutual information is maximized by a Gaussian input whereas lautum information is minimized.

An interesting property of mutual information is the saddle-point characterization of the Gaussian distribution:

$$I(\mathbf{X}; \mathbf{H}\mathbf{X} + \mathbf{N}_G) \leq I(\mathbf{X}_G; \mathbf{H}\mathbf{X}_G + \mathbf{N}_G) \leq I(\mathbf{X}_G; \mathbf{H}\mathbf{X}_G + \mathbf{N}) \quad (45)$$

where  $\mathbf{X}_G$  and  $\mathbf{N}_G$  follow Gaussian distributions and  $\mathbf{X}$  and  $\mathbf{N}$  follow arbitrary distributions with the same first- and second-order moments as the Gaussian counterparts. Unfortunately,

<sup>2</sup>If  $\mathbb{E}[\mathbf{X}] \neq 0$ , then the same result holds substituting  $\mathbf{Y}$  by  $\mathbf{Y} - \mathbf{H}\mathbb{E}[\mathbf{X}]$ .

<sup>3</sup>For the case of real-valued random variables, (42)-(43) require a factor 1/2 in front of the term with the trace.

<sup>4</sup>The proof of  $L(\mathbf{X}; \mathbf{Y}) \geq I(\mathbf{X}; \mathbf{Y})$  for the BSC follows from (is, in fact, equivalent to) the recent refinement of Pinsker's inequality in [15, Th. 2.1] particularized to the binary case.

lautum information does not admit a similar saddle-point characterization. As previously argued, Gaussian inputs minimize lautum information:

$$L(\mathbf{X}_G; \mathbf{H}\mathbf{X}_G + N_G) \leq L(\mathbf{X}; \mathbf{H}\mathbf{X} + N_G). \quad (46)$$

However, the other required inequality for the saddle-point characterization is not satisfied; simply by choosing  $P_N$  such that it vanishes at some set of nonzero measure we obtain  $L(\mathbf{X}_G; \mathbf{H}\mathbf{X}_G + N) = +\infty$ , which implies

$$L(\mathbf{X}_G; \mathbf{H}\mathbf{X}_G + N) > L(\mathbf{X}_G; \mathbf{H}\mathbf{X}_G + N_G). \quad (47)$$

For some other examples, however, the inequality is satisfied in the opposite direction such as with a Laplacian noise (with sufficiently small noise power).

*Theorem 6:* Consider the Gaussian signal model in (41) where  $\mathbf{X}$  is Gaussian. Then, the mutual information and lautum information are given by<sup>5</sup>

$$I(\mathbf{X}; \mathbf{Y}) = \log \det (\mathbf{I} + \Phi^{-1} \mathbf{H} \Sigma \mathbf{H}^\dagger) \quad (48)$$

$$L(\mathbf{X}; \mathbf{Y}) = 2 \operatorname{Tr} \left( \Phi^{-1} \mathbf{H} \Sigma \mathbf{H}^\dagger \right) \log e - \log \det (\mathbf{I} + \Phi^{-1} \mathbf{H} \Sigma \mathbf{H}^\dagger). \quad (49)$$

*Proof:* Particularize Theorem 5 using

$$D(P_Y \| P_N) = -\log \det (\mathbf{I} + \Phi^{-1} \mathbf{H} \Sigma \mathbf{H}^\dagger) + \operatorname{Tr} \left( \Phi^{-1} \mathbf{H} \Sigma \mathbf{H}^\dagger \right) \log e. \quad (50)$$

In the scalar case, (48)-(49) reduce to

$$I(X; Y) = \log \left( 1 + \operatorname{snr} |h|^2 \right) \quad (51)$$

$$L(X; Y) = 2 \operatorname{snr} |h|^2 \log e - \log \left( 1 + \operatorname{snr} |h|^2 \right) \quad (52)$$

where  $\operatorname{snr} = \sigma^2 / \phi^2$ . Interestingly, both measures of information coincide for small SNR:

$$I(X; Y) = L(X; Y) = (\log e) \operatorname{snr} |h|^2 + o(\operatorname{snr}). \quad (53)$$

The following result characterizes the sensitivity of both measures of information.

*Theorem 7 ([16]):* Consider the Gaussian signal model in (41) where  $\mathbf{X}$  is arbitrarily distributed. Then, the gradient of mutual information and lautum information with respect to the channel matrix are given by

$$\nabla_{\mathbf{H}} I(\mathbf{X}; \mathbf{Y}) = (\log e) \Phi^{-1} \mathbf{H} \mathbf{E} \quad (54)$$

$$\nabla_{\mathbf{H}} L(\mathbf{X}; \mathbf{Y}) = (\log e) \Phi^{-1} \mathbf{H} (2\Sigma - \mathbf{E}) \quad (55)$$

where  $\mathbf{E} \triangleq \mathbb{E} \left[ (\mathbf{X} - \mathbb{E}[\mathbf{X} | \mathbf{Y}]) (\mathbf{X} - \mathbb{E}[\mathbf{X} | \mathbf{Y}])^\dagger \right]$  is the MMSE matrix.

<sup>5</sup>For the case of real-valued random variables, (48)-(49) require a factor 1/2 on the right-hand side.

## V. LAUTUM INFORMATION FOR JOINTLY GAUSSIAN RANDOM VARIABLES

This section evaluates the mutual information and the lautum information between two proper complex vector joint Gaussian random variables:  $\mathbf{X} \sim \mathcal{CN}(\mathbf{m}_x, \Sigma_x)$  and  $\mathbf{Y} \sim \mathcal{CN}(\mathbf{m}_y, \Sigma_y)$ .

The mutual information can be easily evaluated as

$$I(\mathbf{X}; \mathbf{Y}) = -\log \det (\mathbf{I} - \Sigma_y^{-1} \Sigma_{yx} \Sigma_x^{-1} \Sigma_{xy}). \quad (56)$$

The lautum information can be similarly evaluated:

*Theorem 8:* Let  $(\mathbf{X}, \mathbf{Y})$  be two vector joint Gaussian random variables with covariance matrix  $\begin{bmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_y \end{bmatrix}$ . Then,

$$L(\mathbf{X}; \mathbf{Y}) = \log \det (\mathbf{I} - \Sigma_y^{-1} \Sigma_{yx} \Sigma_x^{-1} \Sigma_{xy}) + 2 \operatorname{Tr} \left( (\mathbf{I} - \Sigma_y^{-1} \Sigma_{yx} \Sigma_x^{-1} \Sigma_{xy})^{-1} - \mathbf{I} \right) \log e. \quad (57)$$

In the scalar case, the result simplifies to

$$L(X; Y) = \log \left( 1 - |\rho|^2 \right) + 2 \left( \frac{1}{1 - |\rho|^2} - 1 \right) \log e \quad (58)$$

where the normalized covariance is given by

$$\rho = \frac{\mathbb{E}[(X - m_x)(Y - m_y)^*]}{\sigma_x \sigma_y}. \quad (59)$$

From (56) and (57), it can be shown that  $L(\mathbf{X}; \mathbf{Y}) \geq I(\mathbf{X}; \mathbf{Y})$ .

## REFERENCES

- [1] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, pp. 79–86, 1951.
- [2] H. Jeffreys, "An invariant form for the prior probability in estimation problems," *Proc. Roy. Soc. Lon., Ser. A*, vol. 186, pp. 453–461, 1946.
- [3] H. Chernoff, "A measure of asymptotic efficiency for tests of a hypothesis based on a sum of observations," *Annals Math. Statist.*, vol. 23, no. 4, pp. 493–507, 1952.
- [4] I. Csiszár, "The method of types," *IEEE Trans. Inform. Theory*, vol. 44, no. 6, pp. 2505–2523, Oct. 1998.
- [5] S. Verdú, *Multuser Detection*. New York, NY, USA: Cambridge University Press, 1998.
- [6] —, "On channel capacity per unit cost," *IEEE Trans. Inform. Theory*, vol. 36, no. 5, pp. 1019–1030, Sept. 1990.
- [7] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York, NY, USA: Wiley, 1991.
- [8] J. Ziv and M. Zakai, "On functionals satisfying a data-processing theorem," *IEEE Trans. Inform. Theory*, vol. IT-19, no. 3, pp. 275–283, May 1973.
- [9] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic Press, 1981.
- [10] S. Kullback, *Information Theory and Statistics*. New York: Wiley, 1959.
- [11] S. Verdú, "Universal estimation of information measures," in *(Invited Talk) IEEE Information Theory Workshop Rotorua*, Rotorua, New Zealand, Aug. 2005.
- [12] T. S. Han and S. Verdú, "Generalizing the Fano inequality," *IEEE Trans. Inform. Theory*, vol. 40, no. 4, pp. 1247–1251, July 1994.
- [13] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.
- [14] R. E. Blahut, *Principles and Practice of Information Theory*. Addison-Wesley, 1987.
- [15] E. Ordentlich and M. J. Weinberger, "A distribution dependent refinement of Pinsker's inequality," *IEEE Trans. Inform. Theory*, vol. 51, no. 5, pp. 1836–1840, May 2005.
- [16] D. P. Palomar and S. Verdú, "Gradient of mutual information in linear vector Gaussian channels," *IEEE Trans. Inform. Theory*, vol. 52, no. 1, Jan. 2006.