Information Theoretic Proofs of Entropy Power Inequalities

Olivier Rioul, *Member, IEEE*

***Abstract*—While most useful information theoretic inequalities can be deduced from the basic properties of entropy or mutual information, up to now Shannon's entropy power inequality (EPI) is an exception: Existing information theoretic proofs of the EPI hinge on representations of differential entropy using either Fisher information or minimum mean-square error (MMSE), which are derived from de Bruijn's identity. In this paper, we first present an unified view of these proofs, showing that they share two essential ingredients: 1) a data processing argument applied to a covariance-preserving linear transformation; 2) an integration over a path of a continuous Gaussian perturbation. Using these ingredients, we develop a new and brief proof of the EPI through a mutual information inequality, which replaces Stam and Blachman's Fisher information inequality (FII) and an inequality for MMSE by Guo, Shamai, and Verdú used in earlier proofs. The result has the advantage of being very simple in that it relies only on the basic properties of mutual information. These ideas are then generalized to various extended versions of the EPI: Zamir and Feder's generalized EPI for linear transformations of the random variables, Takano and Johnson's EPI for dependent variables, Liu and Viswanath's covariance-constrained EPI, and Costa's concavity inequality for the entropy power.**

***Index Terms*—Data processing inequality, de Bruijn's identity, differential entropy, divergence, entropy power inequality (EPI), Fisher information, Fisher information inequality (FII), minimum mean-square error (MMSE), mutual information, relative entropy.**

I. Introduction

In his 1948 historical paper, Shannon proposed the *entropy power inequality* (EPI) [1, Thm. 15], which asserts that the entropy power of the sum of independent random vectors is at least the sum of their entropy powers; equality holds iff[1] the random vectors are Gaussian with proportional covariances. The EPI is one of the deepest inequalities in information theory, and has a long history. Shannon gave a variational argument [1, App. 6] to show that the entropy of the sum of two independent random vectors of given entropies has a stationary point where the two random vectors are Gaussian with proportional covariance matrices, but this does not exclude the possibility that the stationary point is not a global minimum. Stam [2] credits de Bruijn with a first rigorous proof of the EPI in the case where at most one of the random vectors is *not* Gaussian, using a relationship between differential entropy and Fisher information now known as *de Bruijn's identity*. A general proof of the EPI is given by Stam [2] (see also Blachman [3]), based on a related *Fisher information inequality* (FII). Stam's proof is simplified in [4] and [5]. Meanwhile, Lieb [6] proved the EPI via a strengthened Young's inequality from functional analysis. While Lieb's proof does not use information theoretic arguments, Dembo, Cover, and Thomas [4] showed that it can be recast in a unified proof of the EPI and the Brunn-Minkowski inequality in geometry (see also [7] and [8]), which was included in the textbook by Cover and Thomas [9, Sec. 17.8]. Recently, Guo, Shamai, and Verdú [10] found an integral representation of differential entropy using minimum mean-square error (MMSE), which yields another proof of the EPI [11], [12]. A similar, continuous-time proof via causal MMSE was also proposed by Binia [13]. The original information theoretic proofs (by Stam and Blachman, and by Verdú, Guo, and Shamai) were first given for scalar random variables, and then generalized to the vector case either by induction on the dimension [2], [3] or by extending the required tools [4], [11].

The EPI is used to bound capacity or rate-distortion regions for certain types of channel or source coding schemes, especially to prove converses of coding theorems in the case where optimality cannot be resolved by Fano's inequality alone. Shannon used the EPI as early as his 1948 paper [1] to bound the capacity of non-Gaussian additive noise channels. Other examples include Bergmans' solution [14] to the scalar Gaussian broadcast channel problem, generalized to the multiple-input multiple-output (MIMO) case in [15] and [16]; Leung-Yan Cheong and Hellman's determination of the secrecy capacity of the Gaussian wire-tap channel [17], extended to the multiple access case in [18] and [19]; Costa's solution to the scalar Gaussian interference channel problem [20]; Ozarow's solution to the scalar Gaussian source two-description problem [21], extended to multiple descriptions at high resolution in [22]; and Oohama's determination of the rate-distortion regions for various multiterminal Gaussian source coding schemes [23]–[26]. It is interesting to note that in all the above applications, the EPI is used only in the case where all but one of the random vectors in the sum are Gaussian. The EPI for general independent random variables, as well as the corresponding FII, also find application in blind source separation and deconvolution in the context of independent component analysis (see, e.g., [27]–[29]), and is instrumental in proving a strong version of the central limit theorem with convergence in relative entropy [5], [30]–[35].

It appears that the EPI is perhaps the only useful information theoretic inequality that is *not* proved through basic properties of entropy or mutual information. In this paper, we fill the gap by providing a new proof, with the following nice features:

Manuscript received April 13, 2007; revised August 24, 2010; accepted August 29, 2010. Date of current version December 27, 2010. The material in this paper was presented in part at the IEEE International Symposium on Information Theory, Nice, France, 2007.

The author is with the Institut Télécom, Télécom ParisTech, CNRS LTCI, Paris, France.

Communicated by K. Kobayashi, Associate Editor for Shannon Theory.

Digital Object Identifier 10.1109/TIT.2010.2090193

[1]if and only if.

- it hinges solely on the elementary properties of Shannon's mutual information, sidestepping both Fisher's information and MMSE. Thus, it relies only on the most basic principles of information theory;
- it does not require scalar or vector identities such as de Bruijn's identity, nor integral representations of differential entropy;
- the vector case is handled just as easily as the scalar case, along the same lines of reasoning; and
- it goes with a mutual information inequality (MII), which has its own interest.

Before turning to this proof, we make a detailed analysis of the existing information theoretic proofs[2] of the EPI. The reasons for this presentation are as follows:

- it gives some idea of the level of difficulty that is required to understand conventional proofs. The new proof presented in this paper is comparatively simpler and shorter;
- it focuses on the essential ingredients common to all information theoretic proofs of the EPI, namely data processing inequalities and integration over a path of continuous Gaussian perturbation. This serves as a insightful guide to understand the new proof which uses the same ingredients, though in an more expedient fashion;
- it simplifies some of the conventional argumentation and provides intuitive interpretations for the Fisher information and de Bruijn's identity, which have their own interests and applications. In particular, a new, simple proof of a (generalized) de Bruijn's identity, based on a well-known estimation theoretic relationship between relative entropy and Fisher information, is provided;
- it offers a unified view of the apparently unrelated existing proofs of the EPI. They do not only share essentials, but can also be seen as variants of the same proof; and
- it derives the theoretical tools that are necessary to further discuss the relationship between the various approaches, especially for extended versions of the EPI.

The EPI has been generalized in various ways. Costa [36] (see also [37]) strengthened the EPI for two random vectors in the case where one of these vectors is Gaussian, by showing that the entropy power is a concave function of the power of the added Gaussian noise. Zamir and Feder [38]–[40] generalized the scalar EPI by considering the entropy power of an arbitrary linear transformation of the random variables. Takano [41] and Johnson [42] provided conditions under which the original EPI still holds for two dependent variables. Recently, Liu and Viswanath [43], [44] generalized the EPI by considering a covariance-constrained optimization problem motived by multiterminal coding problems. The ideas in the new proof of the EPI presented in this paper are readily extended to all these situations. Again, in contrast to existing proofs, the obtained proofs rely only on the basic properties of entropy and mutual information. In some cases, further generalizations of the EPI are provided.

The remainder of this paper is organized as follows. We begin with some notations and preliminaries. Section II surveys earlier information theoretic proofs of the EPI and presents a unified view of the proofs. Section III gives the new proof of the EPI, along with some discussions and perspectives. The reader may wish to skip directly to the proof in this section, which does not use the tools presented earlier. Section IV extends the new proof to Zamir and Feder's generalized EPI for arbitrary linear transformations of independent variables. Section V adapts the new proof to the case of dependent random vectors, generalizing the results of Takano and Johnson. Section VI generalizes the new proof to an explicit formulation of Liu and Viswanath's EPI under a covariance constraint, based on the corresponding MII. Section VII gives a proof of the concavity of the entropy power (Costa's EPI) based on the MII, which relies only on the properties of mutual information. Section VIII concludes this paper with some open questions about a recent generalization of the EPI to arbitrary subsets of independent variables [45]–[48] and a collection of convexity inequalities for linear "gas mixtures."

[2]Lieb's excepted, since it belongs to mathematical analysis and cannot be qualified as an "information theoretic" proof.

A. Notations

In this paper, to avoid $\log e$ factors in the derivations, information quantities are measured in *nats*—we shall use only natural logarithms and exponentials. Random variables or vectors are denoted by upper case letters, and their values denoted by lower case letters. The expectation $\mathsf{E}(\cdot)$ is taken over the joint distribution of the random variables within the parentheses. The covariance matrix of a random (column) n-vector X is $\mathsf{Cov}(X) = \mathsf{E}((X-\mathsf{E}(X))(X-\mathsf{E}(X))^t)$, and its variance is the trace of the covariance matrix: $\mathsf{Var}(X) = \mathsf{tr}(\mathsf{Cov}(X)) = \mathsf{E}(\|X-\mathsf{E}(X)\|^2)$. We also use the notation $\sigma_X^2 = \frac{1}{n}\mathsf{Var}(X)$ for the variance per component. We say that X is *white* if its covariance matrix is proportional to the identity matrix, and *standard* if it has unit covariance matrix $\mathsf{Cov}(X) = \mathbf{I}$.

With the exception of the conditional mean $\mathsf{E}(X\,|\,Y)$, which is a function of Y, all quantities in the form $f(X\,|\,Y)$ used in this paper imply expectation over Y, following the usual convention for conditional information quantities. Thus the conditional covariance matrix is $\mathsf{Cov}(X\,|\,Y) = \mathsf{E}((X-\mathsf{E}(X\,|\,Y))(X-\mathsf{E}(X\,|\,Y))^t)$, and the conditional variance is $\mathsf{Var}(X\,|\,Y) = \mathsf{tr}(\mathsf{Cov}(X\,|\,Y)) = \mathsf{E}(\|X-\mathsf{E}(X\,|\,Y)\|^2)$, that is, the MMSE in estimating X given the observation Y, achieved by the conditional mean estimator $\hat{X}(Y) = \mathsf{E}(X\,|\,Y)$.

The diagonal matrix with entries a_i is denoted by $\mathsf{diag}(a_i)_i$. We shall use the partial ordering between real symmetric matrices where $\mathbf{A} \leq \mathbf{B}$ means that the difference is positive semidefinite, that is, for any real vector x, $x^t\mathbf{A}x \leq x^t\mathbf{B}x$. Clearly $\mathbf{A} \leq \mathbf{B}$ implies $\mathbf{CAC} \leq \mathbf{CBC}$ for any symmetric matrix $\mathbf{C}$, and $\mathbf{B}^{-1} \leq \mathbf{A}^{-1}$ if $\mathbf{A}$ and $\mathbf{B}$ are invertible and $\mathbf{A}$ is positive semidefinite.

Given a function $f(x)$, $\frac{\partial f}{\partial x}$ denotes the gradient, a (column) vector of partial derivatives $(\frac{\partial f}{\partial x_i})_i$, and $\frac{\partial^2 f}{\partial x^2}$ denotes the Hessian, a matrix of second partial derivatives $(\frac{\partial^2 f}{\partial x_i \partial x_j})_{i,j}$. We shall use Landau's notations $o(f)$ (a function which is negligible compared to f in the neighborhood of some limit value of x) and $O(f)$ (a function which is dominated by f in that neighborhood).

B. Definition of the Differential Entropy

Let X be any random n-vector having probability density $p(x)$ (with respect to the Lebesgue measure). Its (differential)

entropy is defined by

$$h(X) = \mathsf{E}\log\frac{1}{p(X)} = -\int p(x)\log p(x)\,dx \tag{1}$$

provided that this integral exists in the generalized sense—that is, the positive and negative parts of this integral are not both infinite. Thus we may have $h(X) = +\infty$ if $\mathsf{E}\log^+\frac{1}{p(X)} = +\infty$ and $\mathsf{E}\log^-\frac{1}{p(X)} < +\infty$; and $h(X) = -\infty$ if $\mathsf{E}\log^+\frac{1}{p(X)} < +\infty$ and $\mathsf{E}\log^-\frac{1}{p(X)} = +\infty$, where we have noted $x^+ = \max(x,0)$ and $x^- = \max(-x,0)$.

The differential entropy is not always well defined. Take, for example, $p(x) = \frac{1}{2x\log^2 x}$ for $0 < x < 1/e$ and $e < x < +\infty$, and $p(x) = 0$ otherwise. In this case it is easy to check that both positive and negative parts of the integral $\int p(x)\log p(x)\,dx$ are infinite. In spite of that differential entropy is frequently encountered in the literature, the author was unable to find simple, general conditions under which it is well defined. An exception is [49] which gives the sufficient condition that $p^\alpha(x)$ is Lebesgue-integrable for any α in the range $\alpha_0 \le \alpha \le 2$ where $0 < \alpha_0 < 1$. The following result may be more useful for practical considerations.

Proposition 1 (Well-Defined Entropy): If $\mathsf{E}(\log(1+\|X\|))$ is finite, in particular if X has finite first or second moments, then $h(X)$ is well defined and is such that $-\infty \le h(X) < +\infty$.

Proof: It is sufficient to prove that the positive part $h^+(X) = \mathsf{E}\log^+\frac{1}{p(X)}$ of (1) is finite. Let $q(x)$ be the Cauchy density defined by

$$q(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\pi^{\frac{n+1}{2}}}\frac{1}{(1+\|x\|^2)^{\frac{n+1}{2}}}. \tag{2}$$

Since $u\log u \ge -1/e$ for all $u > 0$, we have

$$\begin{aligned}
h^+(X) &= -\int_{0<p(x)\le 1} p(x)\log p(x)dx && \text{(3a)}\\
&= -\int_{0<p(x)\le 1} p(x)\log q(x)dx \\
&\quad + \int_{0<p(x)\le 1} q(x)\frac{p(x)}{q(x)}\log\frac{q(x)}{p(x)}dx && \text{(3b)}\\
&\le \log\frac{\pi^{\frac{n+1}{2}}}{\Gamma(\frac{n+1}{2})} + \frac{n+1}{2}\mathsf{E}(\log(1+\|X\|^2)) \\
&\quad + \frac{1}{e}\int_{p(x)\le 1} q(x)dx && \text{(3c)}\\
&\le \log\frac{\pi^{\frac{n+1}{2}}}{\Gamma\left(\frac{n+1}{2}\right)} + (n+1)\mathsf{E}(\log(1+\|X\|)) + \frac{1}{e} && \text{(3d)}
\end{aligned}$$

which is finite by assumption. □

It is easy to adapt the proof in the particular case where $\mathsf{E}(\|X\|)$ or $\mathsf{E}(\|X\|^2)$ is finite by letting $q(x)$ be an exponential Laplacian or normal Gaussian distribution, respectively. The proof can also be shorten slightly by applying the theorem of Gel'fand-Yaglom-Perez [50, ch. 2] to the relative entropy $D(p\|q) = \int p(x)\log\frac{p(x)}{q(x)}\,dx$, which is finite because X is absolutely continuous with respect to the measure defined by density $q(x)$.

In the situation of Proposition 1 it is sometimes convenient to extend the definition by setting $h(X) = -\infty$ when X does *not* admit a density with respect to the Lebesgue measure—in particular, when the distribution of X has a probability mass assigned to one or more singletons in $\mathbb{R}^n$ (see, e.g., [36] and [51, p. 6]). This convention can be justified by a limiting argument in several cases. Another justification appears in Lemma 1.

C. EPI

The *entropy power* $N(X)$ of a random n-vector X with (differential) entropy $h(X)$ is [1]

$$N(X) = \frac{e^{\frac{2}{n}h(X)}}{2\pi e}. \tag{4}$$

In the following, we assume that entropies are well defined, possibly with value $h(X) = -\infty$ or $N(X) = 0$. The scaling properties

$$\begin{aligned} h(aX) &= h(X) + n\log|a| \\ N(aX) &= a^2N(X) \end{aligned} \tag{5}$$

where $a \in \mathbb{R}$, follow from the definitions by a change of variable argument.

Suppose X has finite covariances. The *non-Gaussianness* of X is the relative entropy (divergence) with respect to a Gaussian random vector X^* with identical second moments

$$D(X\|X^*) = h(X^*) - h(X) \tag{6}$$

where $h(X^*) = \frac{1}{2}\log((2\pi e)^n|\mathsf{Cov}(X)|)$. With the convention of the preceding section, one has $D(X\|X^*) = +\infty$ if $h(X) = -\infty$. Since (6) is nonnegative and vanishes iff X is Gaussian, the entropy power (4) satisfies

$$N(X) \le |\mathsf{Cov}(X)|^{1/n} \le \sigma_X^2 \tag{7}$$

with equality in the first inequality iff X is Gaussian, and in the second iff X is white. In particular, $N(X)$ is the power of a white Gaussian random vector having the same entropy as X.

From these observations, it is easily found that Shannon's EPI can be given several equivalent forms.

Proposition 2 (Equivalent EPIs): The following inequalities, each stated for finitely many independent random vectors $(X_i)_i$ with finite differential entropies, and real-valued coefficients $(a_i)_i$, are equivalent.

$$N\Big(\sum_i a_iX_i\Big) \ge \sum_i a_i^2N(X_i) \tag{8a}$$

$$h\Big(\sum_i a_iX_i\Big) \ge h\Big(\sum_i a_i\tilde{X}_i\Big) \tag{8b}$$

$$h\Big(\sum_i a_iX_i\Big) \ge \sum_i a_i^2h(X_i) \quad \Big(\sum_i a_i^2 = 1\Big) \tag{8c}$$

where the $(\tilde{X}_i)_i$ are independent Gaussian random vectors with proportional covariances (e.g., white) and corresponding entropies $h(\tilde{X}_i) = h(X_i)$.

We have presented weighted forms of the inequalities to stress the similarity between (8a)–(8c). Note that by (5), the normalization $\sum_i a_i^2 = 1$ is unnecessary for (8a) and (8b). The proof

is given in [4] and is also partly included in [11] in the scalar case. For completeness we include a short proof[3].

Proof: That (8a), (8b) are equivalent follows from the equalities $\sum_i a_i^2 N(X_i) = \sum_i a_i^2 N(\tilde{X}_i) = N(\sum_i a_i \tilde{X}_i)$. To prove that (8c) is equivalent to (8a) we may assume that $\sum_i a_i^2 = 1$. Taking logarithms of both sides of (8a), inequality (8c) follows from the concavity of the logarithm. Conversely, taking exponentials of both sides of (8c), inequality (8a) follows provided that the $(X_i)_i$ have equal entropies. But the latter condition is unnecessary because if (8a) is satisfied for the random vectors $(N(X_i)^{-1/2}X_i)_i$ of equal entropies, then upon modification of the coefficients it is also satisfied for the $(X_i)_i$. □

Inequality (8a) is equivalent to the classical formulation of the EPI [1] by virtue of the scaling property (5). Inequality (8b) is implicit in [1, App. 6], where Shannon's line of thought is to show that the entropy of the sum of independent random vectors of given entropies has a minimum where the random vectors are Gaussian with proportional covariance matrices. It was made explicit by Costa and Cover [7]. Inequality (8c) is due to Lieb [6] and is especially interesting since all available proofs of the EPI are in fact proofs of this inequality. It can be interpreted as a concavity property of entropy [4] under the covariance-preserving transformation

$$(X_i)_i \longmapsto Y = \sum_i a_i X_i \quad \big(\sum_i a_i^2 = 1\big). \tag{9}$$

Interestingly, (8c) is most relevant in several applications of the EPI. Although the preferred form for use in coding applications [14]–[26] is $N(X+Z) \geq N(X) + N(Z)$, where Z is Gaussian independent of X, Liu and Viswanath [43], [44] suggest that the EPI's main contribution to multiterminal coding problems is for solving optimization problems of the form $\max_X h(X) - \mu h(X+Z)$, whose solution is easily determined from the convexity inequality (8c) as shown in Section VI. Also, (8c) is especially important for solving blind source separation and deconvolution problems, because it implies that negentropy $c = -h$ satisfies the requirements for a "contrast function"

$$c\big(\sum_i a_i X_i\big) \leq \max_i c(X_i) \quad \big(\sum_i a_i^2 = 1\big) \tag{10}$$

which serves as an objective function to be maximized in such problems [27]–[29]. Finally, the importance of the EPI for proving strong versions of the central limit theorem is through (8c) interpreted as a monotonicity property of entropy for standardized sums of independent variables [30], [34].

II. Earlier Proofs Revisited

A. Fisher Information Inequalities (FII)

Conventional information theoretic proofs of the EPI use an alternative quantity, the Fisher information (or a disguised version of it), for which the statements corresponding to (8) are easier to prove. The *Fisher information matrix* $\mathbf{J}(X)$ of a random n-vector X with density $p(x)$ is [4], [9]

$$\mathbf{J}(X) = \mathsf{Cov}(S(X)) \tag{11}$$

where the zero-mean random variable (log-derivative of the density)

$$S(X) = \nabla \log p(X) = \frac{\nabla p(X)}{p(X)} \tag{12}$$

is known as the *score*. The *Fisher information* $J(X)$ is the trace of (11)

$$J(X) = \mathsf{Var}(S(X)) = \mathsf{E}\frac{\|\nabla p(X)\|^2}{p(X)^2}. \tag{13}$$

In this and the following subsections, we assume that probability densities are sufficiently smooth with sufficient decay at infinity so that Fisher informations exist, possibly with the value $J(X) = +\infty$. The scaling properties

$$\begin{aligned} S(aX) &= a^{-1}S(X) \\ J(aX) &= a^{-2}J(X) \end{aligned} \tag{14}$$

follow from the definitions by a change of variable argument. Note that if X has independent entries, then $\mathbf{J}(X)$ is the diagonal matrix $\mathbf{J}(X) = \mathsf{diag}(J(X_i))_i$.

It is easily seen that the score $S(X)$ is a linear function of X iff X is Gaussian. Therefore, a measure of *non-Gaussianness* of X is the mean-square error of the score with respect to the (linear) score S^* of a Gaussian random vector X^* with identical second moments:

$$\mathsf{E}(\|S(X) - S^*(X)\|^2) = J(X) - J(X^*) \tag{15}$$

where $J(X^*) = \mathsf{tr}(\mathsf{Cov}(X)^{-1})$. Since (15) is nonnegative and vanishes iff X is Gaussian, the Fisher information (13) satisfies the inequalities

$$J(X) \geq \mathsf{tr}(\mathsf{Cov}(X)^{-1}) \geq \frac{n}{\sigma_X^2}. \tag{16}$$

The first inequality (an instance of the Cramér-Rao inequality) holds with equality iff X is Gaussian, while the second inequality (a particular case of the Cauchy-Schwarz inequality on the eigenvalues of $\mathsf{Cov}(X)$) holds with equality iff X is white. In particular, $nJ^{-1}(X)$ is the power of a white Gaussian random vector having the same Fisher information as X.

Proposition 3 (Equivalent FIIs): The following inequalities, each stated for finitely many independent random vectors $(X_i)_i$ with finite Fisher informations, and real-valued coefficients $(a_i)_i$, are equivalent:

$$J^{-1}\big(\sum_i a_i X_i\big) \geq \sum_i a_i^2 J^{-1}(X_i) \tag{17a}$$

$$J\big(\sum_i a_i X_i\big) \leq J\big(\sum_i a_i \tilde{X}_i\big) \tag{17b}$$

$$J\big(\sum_i a_i X_i\big) \leq \sum_i a_i^2 J(X_i) \quad \big(\sum_i a_i^2 = 1\big) \tag{17c}$$

[3]This proof corrects a small error in [4], namely, that the first statement in the proof of Theorem 7 in [4] is false when the Gaussian random vectors do not have identical covariances.

where the $(\tilde{X}_i)_i$ are independent Gaussian random vectors with proportional covariances (e.g., white) and corresponding Fisher informations $J(\tilde{X}_i) = J(X_i)$.

There is a striking similarity with Proposition 2. The proof is the same, with the appropriate changes—the convexity of the hyperbolic $1/x$ is used in place of the concavity of the logarithm—and is omitted. Inequality (17c) is due by Stam and its equivalence with (17a) was pointed out to him by de Bruijn [2]. It can be shown [52], [53] that the above inequalities also hold for positive semidefinite symmetric matrices, where Fisher informations (13) are replaced by Fisher information matrices (11).

Similarly as for (8c), inequality (17c) can be interpreted as a convexity property of Fisher information [4] under the covariance-preserving transformation (9), or as a monotonicity property for standardized sums of independent variables [30], [35]. It implies that the Fisher information $C = J$ satisfies (10), and therefore, can be used as a contrast function in deconvolution problems [27]. The FII has also been used to prove a strong version of the central limit theorem [5], [30]–[35] and a characterization of the Gaussian distribution by rotation [52], [54].

B. Data Processing Inequalities for Least Squares Estimation[4]

Before turning to the proof the FII, it is convenient and useful to make some preliminaries about data processing inequalities for Fisher information and MMSE. In estimation theory, the importance of the Fisher information follows from the Cramér-Rao bound (CRB) [9] on the mean-squared error of an estimator of a parameter $\theta \in \mathbb{R}^m$ from a measurement $X \in \mathbb{R}^n$. In this context, X is a random n-vector whose density $p_\theta(x)$ depends on θ, and the (parametric) *Fisher information matrix* is defined by [9], [53]

$$\mathbf{J}_\theta(X) = \mathsf{Cov}(S_\theta(X)) \tag{18}$$

where $S_\theta(X)$ is the (parametric) *score function*

$$S_\theta(X) = \frac{\partial}{\partial\theta} \log p_\theta(X). \tag{19}$$

In some references, the parametric Fisher information is defined as the trace of (18)

$$J_\theta(X) = \mathsf{Var}(S_\theta(X)). \tag{20}$$

In the special case where $\theta \in \mathbb{R}^n$ is a *translation parameter*: $p_\theta(x) = p(x+\theta)$, we recover the earlier definitions (11)–(13): $S(X) = S_\theta(X-\theta)$, $\mathbf{J}(X) = \mathbf{J}_\theta(X-\theta)$, and $J(X) = J_\theta(X-\theta)$. More generally, it is easily checked that for any $a \in \mathbb{R}$

$$S_\theta(X - a\theta) = aS(X) \tag{21a}$$
$$J_\theta(X - a\theta) = a^2 J(X). \tag{21b}$$

The optimal unbiased estimator of θ given the observation X, if it exists, is such that the mean-square error meets the CRB (reciprocal of the Fisher information) [9]. Such an optimal estimator is easily seen to be a linear function of the score (19). Thus it may be said that the score function $S_\theta(X)$ represents the optimal least squares estimator of θ. When the estimated quantity θ is a random variable (i.e., not a parameter), the optimal estimator is the conditional mean estimator $\mathsf{E}(\theta \,|\, X)$ and the corresponding minimum mean-square error (MMSE) is the conditional variance $\mathsf{Var}(\theta \,|\, X)$.

In both cases, there is a *data processing theorem* [53] relative to a transformation $X \to Y$ in a *Markov chain* $\theta \to X \to Y$, that is, for which Y given X is independent of θ. The emphasize the similarity between these data processing theorems and the corresponding quantities of Fisher information and MMSE, we first prove the following "chain rule," which states that the optimal estimation given Y of θ results from the optimal estimation given Y of the optimal estimation given X of θ.

Proposition 4 (Data Processing Theorem for Estimators): If $\theta \to X \to Y$ form a Markov chain, then

$$\mathsf{E}(\theta \,|\, Y) = \mathsf{E}(\mathsf{E}(\theta \,|\, X) \,|\, Y) \tag{22a}$$
$$S_\theta(Y) = \mathsf{E}_\theta(S_\theta(X) \,|\, Y). \tag{22b}$$

Proof: In the nonparametric case the Markov chain condition can written as $p(\theta \,|\, x, y) = p(\theta \,|\, x)$. Multiplying by $\theta\, p(x \,|\, y)$ gives $\theta\, p(\theta, x \,|\, y) = \theta\, p(\theta \,|\, x) p(x \,|\, y)$, which integrating over θ and x yields (22a). In the parametric case the Markov chain condition can be written as $p_\theta(x, y) = p_\theta(x) p(y \,|\, x)$ where the distribution $p(y \,|\, x)$ is independent of θ. Differentiating with respect to θ gives $\frac{\partial p_\theta}{\partial\theta}(x, y) = \frac{\partial p_\theta}{\partial\theta}(x) p(y \,|\, x)$; dividing by $p_\theta(y)$ and applying Bayes' rule yields $\frac{\partial p_\theta}{\partial\theta}(x, y)/p_\theta(y) = \frac{\partial p_\theta}{\partial\theta}(x)/p_\theta(x)\, p_\theta(x \,|\, y)$, which integrating over x yields (22b). □

From Proposition 4 we obtain a unified proof of the corresponding *data processing inequalities* for least squares estimation, which assert that the transformation $X \to Y$ reduces information about θ, or in other words, that no clever transformation can improve the inferences made on the data measurements: compared to X, the observation Y yields a worse estimation of θ.

Proposition 5 (Estimation Theoretic Data Processing Inequalities): If $\theta \to X \to Y$ form a Markov chain, then

$$\mathsf{Cov}(\theta \,|\, Y) \geq \mathsf{Cov}(\theta \,|\, X) \tag{23a}$$
$$\mathbf{J}_\theta(Y) \leq \mathbf{J}_\theta(X). \tag{23b}$$

In particular

$$\mathsf{Var}(\theta \,|\, Y) \geq \mathsf{Var}(\theta \,|\, X) \tag{24a}$$
$$J_\theta(Y) \leq J_\theta(X). \tag{24b}$$

Equality holds iff

$$\mathsf{E}(\theta \,|\, X) = \mathsf{E}(\theta \,|\, Y) \text{ a.e.}, \tag{25a}$$
$$S_\theta(X) = S_\theta(Y) \text{ a.e.} \tag{25b}$$

respectively.

Proof: The following identity ("total law of covariance") is well known and easy to check:

$$\mathsf{Cov}(U) = \mathsf{Cov}(U \,|\, V) + \mathsf{Cov}(\mathsf{E}(U \,|\, V)). \tag{26}$$

[4]We use the term "least squares estimation" for any estimation procedure based on the mean-squared error criterion.

For $U = \mathsf{E}(\theta \,|\, X)$ or $U = S_\theta(X)$, and $V = Y$, we obtain, by Proposition 4

$$\mathsf{Cov}(\theta \,|\, X) = \mathsf{Cov}(\theta \,|\, Y) - \mathsf{Cov}(\mathsf{E}(\theta \,|\, X) \,|\, Y) \quad (27a)$$

$$\mathbf{J}_\theta(X) = \mathbf{J}_\theta(Y) + \mathsf{Cov}(S_\theta(X) \,|\, Y) \quad (27b)$$

where in deriving (27a) we have also used (26) for $U = \theta$. Since covariance matrices are positive semidefinite, this proves (23a), (23b), and (24a), (24b) follow by taking the trace. Equality holds in (23a), (24a), or in (23b), (24b) iff $E(\theta \,|\, X)$ or $S_\theta(X)$ is a deterministic function of Y, which by Proposition 4 is equivalent to (25a) or (25b), respectively. □

Stam [2] mentioned that (24b) is included in the original work of Fisher, in the case where Y is a deterministic function of X. A different proof of (23b) is provided by Zamir [53]. The above proof also gives, via (27a), (27b) or the corresponding identities for the variance, explicit expressions for the information "loss" due to processing. The equality conditions correspond to the case where the optimal estimators given X or Y are the same. In particular, it is easily checked that (25b) is equivalent to the fact that $\theta \to Y \to X$ (in this order) also form a Markov chain, that is, Y is a "sufficient statistic" relative to X [9].

As a consequence of Proposition 5, we obtain a simple proof of the following relation between Fisher information and MMSE in the case where estimation is made in Gaussian noise.

Proposition 6 (Complementary Relation between Fisher Information and MMSE): If Z is Gaussian independent of X, then

$$\mathbf{J}(X+Z)\mathsf{Cov}(Z) + \mathsf{Cov}(Z)^{-1}\mathsf{Cov}(X \,|\, X+Z) = \mathbf{I}. \quad (28)$$

In particular, if Z is white Gaussian

$$\sigma_Z^2 J(X+Z) + \sigma_Z^{-2}\mathsf{Var}(X \,|\, X+Z) = n. \quad (29)$$

Proof: Apply (27b) to the Markov chain $\theta \to (X, Z-\theta) \to X+Z-\theta$, where X and Z are independent of θ and of each other. Since $S_\theta(X, Z-\theta) = S_\theta(X) + S_\theta(Z-\theta) = S(Z) = -\mathsf{Cov}(Z)^{-1}(Z - \mathsf{E}(Z))$, we have $J_\theta(X, Z-\theta) = J(Z) = \mathsf{Cov}(Z)^{-1}$. Therefore, (27b) reads $\mathsf{Cov}(Z)^{-1} = \mathbf{J}(X+Z) + \mathsf{Cov}(Z)^{-1} \times \mathsf{Cov}(Z \,|\, X+Z)\mathsf{Cov}(Z)^{-1}$. Noting that $Z - \mathsf{E}(Z \,|\, X+Z) = \mathsf{E}(X \,|\, X+Z) - X$, one has $\mathsf{Cov}(Z \,|\, X+Z) = \mathsf{Cov}(X \,|\, X+Z)$ and (28) follows upon multiplication by $\mathsf{Cov}(Z)$. For white Gaussian Z, (29) follows by taking the trace. □

As noted by Madiman and Barron [47], (29) is known in Bayesian estimation (average risk optimality): see [55, Th. 4.3.5] in the general situation where $X+Z$ is replaced any variable Y such that $p(y \,|\, x)$ belongs to an exponential family parameterized by x. It was rediscovered independently by Budianu and Tong [56], and by Guo, Shamai, and Verdú [10], [57]. Relation (28) was also rederived by Palomar and Verdú [58] as a consequence of a generalized de Bruijn's identity (Corollary 1). Other existing proofs are by direct calculation. The above proof is simpler and offers an intuitive alternative based on the data processing theorem.

To illustrate (28), consider the case where X and Z are zero-mean Gaussian. In this case, the conditional mean estimator $\mathsf{E}(X \,|\, X+Z)$ is linear of the form $\mathbf{A}(X+Z)$, where $\mathbf{A}$ is given by the Wiener-Hopf equations $\mathbf{A}\mathsf{Cov}(X+Z) = \mathsf{E}(X(X+Z)^t) = \mathsf{Cov}(X)$. Therefore, $\mathsf{E}(X \,|\, X+Z) = \mathsf{Cov}(X)\mathsf{Cov}(X+Z)^{-1}(X+Z) = X + Z - \mathsf{Cov}(Z)\mathsf{Cov}(X+Z)^{-1}(X+Z)$. This gives, after some calculations, $\mathsf{Cov}(X \,|\, X+Z) = \mathsf{Cov}(Z) - \mathsf{Cov}(Z)\mathsf{Cov}(X+Z)^{-1}\mathsf{Cov}(Z)$. But this expression is also an immediate consequence of (28) since one has simply $\mathbf{J}(X+Z) = \mathsf{Cov}(X+Z)^{-1}$.

For standard Gaussian Z, (29) reduces to the identity $J(X+Z) + \mathsf{Var}(X \,|\, X+Z) = n$, which constitutes a simple complementary relation between Fisher information and MMSE. The estimation of X from the noisy version $X+Z$ is all the more better as the MMSE is lower, that is, as $X+Z$ has higher Fisher information. Thus Fisher information can be interpreted a measure of least squares (nonparametric) estimation's efficiency, when estimation is made in additive Gaussian noise.

C. Proofs of the FII via Data Processing Inequalities

Three distinct proofs of the FII (17c) are available in the literature. In this section, we show that these are in fact variations on the same theme: thanks to the presentation of Section II-B, each proof can be easily interpreted as an application of the data processing theorem to the (linear) deterministic transformation $(X_i)_i \mapsto Y$ given by (9), or in parametric form

$$Y - \theta = \sum_i a_i(X_i - a_i\theta) \quad \big(\sum_i a_i^2 = 1\big). \quad (30)$$

1) Proof via the Data Processing Inequality for Fisher Information: This is essentially Stam's proof [2] (see also Zamir [53] for a direct proof of (17a) by this method). Simply apply (24b) to the transformation (30)

$$J_\theta\big(\sum_i a_i X_i - \theta\big) \le J_\theta((X_i - a_i\theta)_i) = \sum_i J_\theta(X_i - a_i\theta). \quad (31)$$

From (21b), the FII (17c) follows.

2) Proof via Conditional Mean Representations of the Score: This proof is due to Blachman [3] in the scalar case ($n = 1$). His original derivation is rather technical, since it involves a direct calculation of the convolution of the densities of independent random variables U and V to establish that $S(U+V) = \mathsf{E}(\lambda S(U) + (1-\lambda)S(V) \,|\, U+V)$ for any $0 \le \lambda \le 1$, followed by an application of the Cauchy-Schwarz inequality. The following derivation is simpler and relies on the data processing theorem: By Proposition 4 applied to (30)

$$\begin{aligned} S_\theta\big(\sum_i a_i X_i - \theta\big) &= \mathsf{E}\big(S_\theta\big((X_i - a_i\theta)_i\big) \,|\, \sum_i a_i X_i - \theta\big) \\ &= \mathsf{E}\big(\sum_i S_\theta(X_i - a_i\theta) \,|\, \sum_i a_i X_i\big) \end{aligned}$$

which from (21a) gives the following conditional mean representation of the score:

$$S\big(\sum_i a_i X_i\big) = \mathsf{E}\big(\sum_i a_i S(X_i) \,|\, \sum_i a_i X_i\big). \quad (32)$$

This representation includes Blachman's as a special case (for two variables $U = a_1X_1$ and $V = a_2X_2$). The rest of Blachman's argument parallels the above proof of the data processing inequality for Fisher information (Proposition 5): His application of the Cauchy-Schwarz inequality [3] is simply a consequence of the law of total variance

$\mathsf{Var}(U) = \mathsf{Var}(U\,|\,V) + \mathsf{Var}(\mathsf{E}(U\,|\,V))$. Indeed, taking $U = \sum_i a_i S(X_i)$, $V = \sum_i a_i X_i$, and using (32), the inequality $\mathsf{Var}(U) \geq \mathsf{Var}(\mathsf{E}(U\,|\,V))$ reduces to the FII (17c). Thus we see that, despite appearances, the above two proofs of Stam and Blachman are completely equivalent.

3) Proof via the Data Processing Inequality for MMSE: This proof is due to Verdú and Guo [11], which use MMSE in lieu of Fisher's information. Apply (24a) to the transformation (9), in which each X_i is replaced by $X_i + Z_i$, where the $(Z_i)_i$ are i.i.d. white Gaussian of variance σ^2. Noting $Z = \sum_i a_i Z_i$, this gives

$$\begin{aligned}\mathsf{Var}\big(\sum_i a_i X_i \,|\, \sum_i a_i X_i + Z\big) &\geq \mathsf{Var}\big(\sum_i a_i X_i \,|\, (X_i + Z_i)_i\big) \\ &= \sum_i a_i^2 \mathsf{Var}(X_i \,|\, X_i + Z_i) \quad (33)\end{aligned}$$

where Z is also white Gaussian of variance σ^2. By the complementary relation (29) (Proposition 6), this inequality is equivalent to the FII $J(\sum_i a_i X_i + Z) \leq \sum_i a_i^2 J(X_i + Z_i)$ and letting $\sigma^2 \to 0$ gives (17c)[5]. Again this proof is equivalent to the preceding ones, by virtue of the complementary relation between Fisher information and MMSE.

4) Conditions for Equality in the FII: The case of equality in (17c) was settled by Stam [2] and Blachman [3]. In Stam's approach, by Proposition 5, (25b), equality holds in (31) iff $\sum_i S_\theta(X_i - a_i\theta) = S_\theta(\sum_i a_i X_i - \theta)$, that is, using (21a)

$$\sum_i a_i S(X_i) = S\big(\sum_i a_i X_i\big) \quad \text{a.e.} \qquad (34)$$

This equality condition is likewise readily obtained in Blachman's approach above. Obviously, it is satisfied only if all scores for which $a_i \neq 0$ are *linear* functions, which means that equality holds in the FII only if the corresponding random vectors are *Gaussian*. In addition, replacing the scores by their expressions for Gaussian random n-vectors in (34), it follows easily by identification that these random vectors have identical covariance matrices. Thus equality holds in (17c) iff all random vectors X_i such that $a_i \neq 0$ are Gaussian with identical covariances.

Verdú and Guo do not derive the case of equality in [11]. From the preceding remarks, however, it follows that equality holds in (33) only if the $(X_i + Z_i)_i$ for which $a_i \neq 0$ are Gaussian—and, therefore, the corresponding $(X_i)_i$ are themselves Gaussian. This result is not evident from estimation-theoretic properties alone in view of the equality condition (25a) in the data processing inequality for the MMSE.

D. De Bruijn's Identity

1) Background: De Bruijn's identity is the fundamental relation between differential entropy and Fisher information, and as such, is used to prove the EPI (8c) from the corresponding FII (17c). This identity can be stated in the form [4]

$$\frac{d}{dt} h(X + \sqrt{t}\,Z)\bigg|_{t=0} = \frac{1}{2} J(X) \qquad (35)$$

where Z is standard Gaussian, independent of the random n-vector X. It is proved in the scalar case in [2], generalized to the vector case by Costa and Cover [7] and to nonstandard Gaussian Z by Johnson and Suhov [33], [42]. The conventional, technical proof of de Bruijn's identity relies on a diffusion equation satisfied by the Gaussian distribution and is obtained by integrating by parts in the scalar case and invoking Green's identity in the vector case. We shall give a simpler and more intuitive proof of a generalized identity for arbitrary (not necessarily Gaussian) Z.

Proposition 7 (De Bruijn's Identity): For any two independent random n-vectors X and Z such that $\mathbf{J}(X)$ exists and Z has finite covariances

$$\frac{d}{dt} h(X + \sqrt{t}\,Z)\bigg|_{t=0} = \frac{1}{2}\,\mathsf{tr}(\mathbf{J}(X)\,\mathsf{Cov}(Z)). \qquad (36a)$$

In particular, if Z is white or X has i.i.d. entries

$$\frac{d}{dt} h(X + \sqrt{t}\,Z)\bigg|_{t=0} = \frac{1}{2}\sigma_Z^2\, J(X). \qquad (36b)$$

2) A Simple Proof of De Bruijn's Identity: The proof is based on the following observation. Setting $\theta = \sqrt{t}$, (36a) can be rewritten as a first-order Taylor expansion in θ^2

$$\begin{aligned}&h(X + \theta Z) - h(X) \\ &\quad = \frac{\theta^2}{2}\,\mathsf{E}((Z - \mathsf{E}(Z))^t \mathbf{J}(X)(Z - \mathsf{E}(Z))) + o(\theta^2). \quad (37)\end{aligned}$$

Now, there is a well known, similar expansion of *relative entropy* (divergence)

$$D_X(p_\theta \| p_{\theta'}) = \mathsf{E}_\theta \log \frac{p_\theta(X)}{p_{\theta'}(X)} \qquad (38)$$

in terms of *parametric* Fisher information (18), for a parameterized family of densities $p_\theta(x)$, $\theta \in \mathbb{R}^m$. Indeed, since the divergence is nonnegative and vanishes for $\theta' = \theta$, its second-order Taylor expansion takes the form [60]

$$D_X(p_\theta \,\|\, p_{\theta'}) = \frac{1}{2}\,(\theta' - \theta)^t \mathbf{J}_\theta(X)(\theta' - \theta) + o(\|\theta' - \theta\|^2) \qquad (39)$$

where $\mathbf{J}_\theta(X)$ is the positive semidefinite Hessian matrix of the divergence, that is, $\mathbf{J}_\theta(X) = \frac{\partial^2}{\partial \theta'^2} D_X(p_\theta \| p_{\theta'})|_{\theta'=\theta} = \mathsf{E}_\theta \frac{\partial^2}{\partial \theta^2} \log \frac{1}{p_\theta(X)}$, which is easily seen to coincide with definition (18)[6]. In view of the similarity between (37) and (39), the following proof of de Bruijn's identity is almost immediate.

Proof of Proposition 7: Let $Y = X + \theta Z$ and write mutual information $I(X + \theta\,Z; Z) = h(X + \theta\,Z) - h(X)$ as a conditional divergence: $I(Y, Z) = D(p(y\,|\,z)\|p(y)) = \mathsf{E}(D(p_X(y - \theta Z)\|p_Y(y))$. Making the change of variable $u = y - \theta z$ gives $I(X + \theta Z; Z) = \mathsf{E}_Z(D(q_0\|q_\theta))$, where $q_\theta(u) = p_{X+\theta Z}(u + \theta z)$ is the parameterized family of densities of a random variable U, and $q_0(u) = p_X(u)$. Therefore, by (39) for scalar θ

$$I(X + \theta Z; Z) = \frac{\theta^2}{2}\mathsf{E}_Z(J_0(U)) + o(\theta^2) \qquad (40)$$

where $J_0(U)$ is the parametric Fisher information of U about $\theta = 0$, which is easily determined as follows.

[5]This continuity argument is justified in [59].

[6]Even though the divergence is not symmetric in (θ, θ'), it is locally symmetric in the sense that (39) is also the second-order Taylor expansion for $D_X(p_{\theta'}\|p_\theta)$.

Expanding $p(y\,|\,z) = p_X(y-\theta z)$ about $\theta = 0$ gives $p(y\,|\,z) = p_X(y) - \theta z^t \nabla p_X(y) + o(\theta)$, and, therefore, $p(y) = \mathsf{E}(p(y\,|\,Z)) = p_X(y) - \theta\,\mathsf{E}(Z)^t \nabla p_X(y) + o(\theta)$, where the limit for $\theta \to 0$ and the expectation have been exchanged, due to Lebesgue's convergence theorem and the fact that Z has finite covariances. It follows that $q_\theta(u) = p_Y(u+\theta z) = q_0(u) + \theta\,(z - \mathsf{E}(Z))^t \nabla p_X(u) + o(\theta)$ so that the (parametric) score of U for $\theta = 0$ is $S_0(U) = \frac{\partial}{\partial\theta} \log q_\theta(U)\,|_{\theta=0} = (z - \mathsf{E}(Z))^t \frac{\nabla p_X(U)}{p_X(U)}$ where $\frac{\nabla p_X}{p_X}$ is the (nonparametric) score of X. Therefore, $J_0(U) = \mathsf{Var}(S_0(U)) = (z - \mathsf{E}(Z))^t \mathbf{J}(X)(z - \mathsf{E}(Z))$. Plugging this expression into (40) gives (37) as required. □

In exploiting the parallelism between (37) and (39), this proof explains the presence of the $1/2$ factor in de Bruijn's identity: this is merely a second-order Taylor expansion factor due to the definition of Fisher information as the second derivative of divergence. Besides, it is mentioned in [4] that (35) holds for any random vector Z whose first four moments coincide with those of the standard Gaussian; here we see that it is sufficient that this condition hold for the second centered moments ($\mathsf{Cov}(Z) = \mathbf{I}$). Also note that it is not required that Z have a density. Thus, (36) also holds for a discrete valued perturbation Z.

3) The Gaussian Case: When Z is Gaussian, de Bruijn's identity (36) is readily extended to positive values of t. Simply substitute $X + \sqrt{t'}\,Z'$ for X, where Z' is independent of Z with the same distribution. By the stability property of the Gaussian distribution under convolution, $X + \sqrt{t'}\,Z' + \sqrt{t}\,Z$ and $X + \sqrt{t+t'}\,Z$ are identically distributed, and, therefore

$$\frac{d}{dt} h(X + \sqrt{t}\,Z) = \frac{1}{2}\,\mathsf{tr}(\mathbf{J}(X + \sqrt{t}\,Z)\,\mathsf{Cov}(Z)). \tag{41a}$$

For white Z, this reduces to

$$\frac{d}{dt} h(X + \sqrt{t}\,Z) = \frac{1}{2}\sigma_Z^2\, J(X + \sqrt{t}\,Z). \tag{41b}$$

Such a generalization cannot be established for non-Gaussian Z, because the Gaussian distribution is the only stable distribution with finite covariances. Using the complementary relation (28) of Proposition 6 and making the change of variable $t' = 1/t$, it is a simple matter of algebra to show that (41a) is equivalent to

$$\frac{d}{dt} h(\sqrt{t}\,X + Z) = \frac{1}{2}\,\mathsf{tr}(\mathsf{Cov}(Z)^{-1}\,\mathsf{Cov}(X\,|\,\sqrt{t}\,X + Z)). \tag{42a}$$

Since $\mathsf{Cov}(Z)^{-1} = \mathbf{J}(Z)$, this alternative identity also generalizes (36a) (with X and Z interchanged). For white Z, it reduces to

$$\frac{d}{dt} h(\sqrt{t}\,X + Z) = \frac{1}{2\sigma_Z^2}\,\mathsf{Var}(X\,|\,\sqrt{t}\,X + Z). \tag{42b}$$

The latter two identities were thoroughly investigated by Guo, Shamai, and Verdú [10]. The above proof, via de Bruijn's identity and Kullback's expansion (39), is shorter than the proofs given in [10], and also has an intuitive interpretation, as shown next.

4) Intuitive Interpretations: Expansions (37) and (39) can be given similar interpretations. In (39), $D_X(p_\theta \| p_{\theta'})$ has local parabolic behavior at vertex $\theta = \theta'$ with curvature $= J_\theta(X)$, which means that for a given (small) value of divergence, θ is known all the more precisely as Fisher information $J_\theta(X)$ is large (see Fig. 1). This confirms that $J_\theta(X)$ is a quantity of "information" about θ. Similarly, (37) shows that the mutual information $I(X + \theta Z; Z)$ between the noisy version $X + \theta\,Z$ of X and the noise Z, seen as a function of the noise amplitude, is locally parabolic about $\theta = 0$ with curvature $= J(X)$. Hence for a given (small) value of noise amplitude θ_0, the noisy variable is all the more dependent on the noise as $J(X)$ is higher (see Fig. 2). Therefore, de Bruijn's identity merely states that Fisher information measures the *sensitivity* to an arbitrary additive independent noise, in the sense that a highly "sensitive" variable, perturbed by a small additive noise, becomes rapidly noise-dependent as the amplitude of the noise increases. This measure of sensitivity of X depends the noise covariances but is independent of the shape of the noise distribution otherwise, due to the fact that de Bruijn's identity remains true for non-Gaussian Z. Also, by the Cramér-Rao inequality (16), a Gaussian variable X^* has lowest sensitivity to an arbitrary additive noise Z. Thus the saddlepoint property of mutual information $I(X + Z; Z) \geq I(X^* + Z; Z)$, classically established for Gaussian Z [9], [61], [62] (see also Proposition 8), is seen to hold to the first order of σ_Z^2 for an arbitrary additive noise Z.

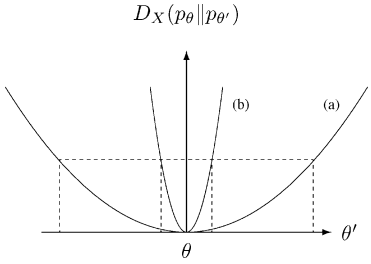

Fig. 1. Kullback-Leibler divergence drawn as a function of the estimated parameter for (a) low and (b) high value of Fisher information.

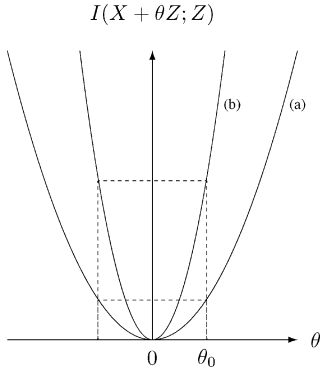

Fig. 2. Mutual information between a noisy variable and the noise, drawn as a function of noise amplitude θ for (a) low and (b) high value of the variable's Fisher information.

A dual interpretation is obtained by exchanging the roles of X and Z in (36a) or (37) to obtain an asymptotic formula for the input-output mutual information $I(X; \sqrt{t}\,X + Z)$ in a (non-Gaussian) additive noise channel $X \mapsto \sqrt{t}\,X + Z$ for small signal-to-noise ratio (SNR). In particular, for i.i.d. input entries or if the channel is memoryless, either $\mathsf{Cov}(X)$ or $\mathbf{J}(Z)$ is proportional to the identity matrix and, therefore

$$I(X; \sqrt{t}\,X + Z) = \frac{1}{2} J(Z)\,\sigma_X^2 t + o(t). \tag{43}$$

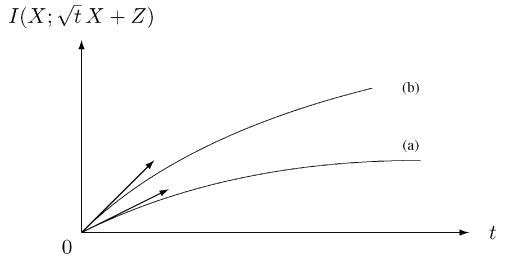

Fig. 3. Input-output mutual information over an additive noise channel, drawn as a function of SNR for small SNR and standard Z. (a) Gaussian channel $J(Z)=1$. (b) Laplacian channel $J(Z)=2$.

Thus, as has been observed in, e.g., [10], [63], and [64], the rate of increase of mutual information per unit SNR is equal to $\frac{1}{2}J(Z)$ in the vicinity of zero SNR, regardless of the shape of the input distribution (see Fig. 3). In the case of a memoryless channel, it is also insensitive to input memory, since in this case (43) still holds for correlated inputs. Again by the Cramér-Rao inequality (16), the Gaussian channel exhibits a minimal rate of increase of mutual information, which complies with the well-known fact that non-Gaussian additive noise channels cannot have smaller capacity than that of the Gaussian channel.

5) Applications: Apart from its role in proving the EPI, de Bruijn's identity (Proposition 7) has found many applications in the literature, although they were not always recognized as such. The Taylor expansion for non-Gaussianness corresponding to (37) in the scalar case ($n=1$) is mentioned, albeit in a disguised form, by Linnik [65] who used it to prove the central limit theorem. Itoh [66] used Linnik's expansion to characterize the Gaussian distribution by rotation. Similar expansions have been derived by Prelov and others (see, e.g., [59], [67]–[78]) to investigate the behavior of the capacity or mutual information in additive Gaussian or non-Gaussian noise channels under various asymptotic scenarios. In particular, (43) was apparently first stated explicitly by Pinsker, Prelov and van der Meulen [59]. A similar result was previously published by Verdú [79] (see also [80]) who used Kullback's expansion (39) to lower bound the capacity per unit SNR for non-Gaussian memoryless additive noise channels, a result which is also an easy consequence of (43). Motivated by the blind source separation problem, Pham [81] (see also [82] and [83]) investigated the first- and second-order expansions in θ of entropy for non-Gaussian perturbation Z (not necessarily independent of X) and recovers de Bruijn's identity as a special case. Similar first- and second-order expansions for mutual information in non-Gaussian additive noise channels were derived by Guo, Shamai, and Verdú [64], yielding (43) as a special case.

6) Generalized De Bruijn's Identity: Palomar and Verdú [58] proposed a matrix version of de Bruijn's identity by considering the gradient of $h(X+Z)$ with respect to the noise covariance matrix $\mathsf{Cov}(Z)$ for Gaussian Z. We call attention that this is a simple consequence of (36a); the generalization to non-Gaussian Z is as follows.

Corollary 1:

$$\frac{d}{d\mathbf{K}}h(X+Z)\bigg|_{\mathbf{K}=\mathbf{0}}=\frac{1}{2}\mathbf{J}(X) \tag{44}$$

where we have noted $\mathbf{K}=\mathsf{Cov}(Z)$.

Proof[7]*:* By (36a), we have the following expansion:

$$h(X+Z)=\frac{1}{2}\,\mathsf{tr}(\mathbf{J}(X)\,\mathbf{K})+o(\|\mathbf{K}\|)$$

where $\|\mathbf{K}\|$ denotes the Fröbenius norm of $\mathbf{K}=\mathbf{K}^t$. But this is of the form of a first-order Taylor expansion of a function with respect to a matrix[8]:

$$f(\mathbf{K})=f(\mathbf{0})+\mathsf{tr}\left(\frac{df}{d\mathbf{K}}(\mathbf{0})\cdot\mathbf{K}^t\right)+o(\|\mathbf{K}\|)$$

and (44) follows by identifying the gradient matrix. □

7) Relationship Between the Cramér-Rao Inequality and a Saddlepoint Property of Mutual Information: The following saddle point property of mutual information, which was proved in [84] using a result of Pinsker [85], states that the worst possible noise distribution in a additive noise channel is the Gaussian distribution.

Proposition 8: Let X be any random vector, and let X^* be a Gaussian random vector with identical second moments. For any Gaussian random vector Z independent of X and X^*

$$I(X+Z;Z)\geq I(X^*+Z;Z). \tag{45}$$

Proof (Following [62]): Noting that $Y^*=X^*+Z$ has identical second moments as $Y=X+Z$, we have $I(X+Z;Z)-I(X^*+Z;Z)=h(Y)-h(X)-h(Y^*)+h(X^*)=D(X\|X^*)-D(Y\|Y^*)$. The result follows by the data processing inequality for divergence, applied to the transformation $X\to Y=X+Z$. □

This proof, in contrast with that given in [9] and [61] for scalar variables, does not require the EPI, and is through a much less involved argument.

Interestingly, by virtue of de Bruijn's identity, it can be shown that (45) is equivalent to the famous Cramér-Rao inequality[9]

$$\mathbf{J}(X)\geq\mathbf{J}(X^*)=\mathsf{Cov}(X)^{-1}. \tag{46}$$

To see this, divide both sides of (45) by the entries of $\mathsf{Cov}(Z)$ and let $\mathsf{Cov}(Z)\to\mathbf{0}$. By Corollary 1, this gives $\frac{1}{2}\mathbf{J}(X)\geq\frac{1}{2}\mathbf{J}(X^*)$. Conversely, integrating the relation $\frac{1}{2}\mathsf{tr}(\mathbf{J}(X+Z)\mathsf{Cov}(Z))\geq\frac{1}{2}\mathsf{tr}(\mathbf{J}(X^*+Z)\mathsf{Cov}(Z))$ using de Bruijn's identity (41) readily gives (45).

E. Earlier Proofs of the EPI

All available information theoretic proofs of the EPI use de Bruijn's identity to integrate the FII (or the corresponding inequality for MMSE) over the path of a continuous Gaussian perturbation. To simplify the presentation, we first consider a path of the form $\{X+\sqrt{t}\,Z\}_{t\in(0;+\infty[}$ where Z is assumed *standard* Gaussian. The derivations in this section are readily extended to the case where Z is arbitrary Gaussian, by means of the corresponding generalized FII and de Bruijn's identity.

[7]The $1/2$ factor is absent in [58], due to the fact that complex gradients are considered.

[8]Putting the matrix entries into a column vector $\mathbf{k}$ it is easily found that $\mathsf{tr}(\frac{df}{d\mathbf{K}}(\mathbf{0})\cdot\mathbf{K}^t)=\mathbf{k}^t\frac{df}{d\mathbf{k}}(\mathbf{0})$.

[9]This follows from the relation $\mathbf{J}(X)-\mathbf{J}(X^*)=\mathsf{Cov}(S(X)-S^*(X))\geq 0$, where $S^*(X)$ is defined as in (15).

1) Basic Proof: The following is a simplified version of Stam's proof [2]. Apply the FII (17c) to the random vectors $(X_i + \sqrt{t}\,Z_i)_i$, where the $(Z_i)_i$ are independent and standard Gaussian. This gives $J(\sum_i a_i X_i + \sqrt{t}\,Z) - \sum_i a_i^2 J(X_i + \sqrt{t}\,Z_i) \le 0$, where $Z = \sum_i a_i Z_i$ is also standard Gaussian. By de Bruijn's identity (41b), it follows that $f(t) = h(\sum_i a_i X_i + \sqrt{t}\,Z) - \sum_i a_i^2 h(X_i + \sqrt{t}\,Z_i)$ is a nonincreasing function of t. But $f(t) = h(t^{-1/2}\sum_i a_i X_i + Z) - \sum_i a_i^2 h(t^{-1/2}X_i + Z_i)$ tends to $h(Z) - \sum_i a_i^2 h(Z_i) = 0$ as $t \to \infty$ (see Lemma 3 below). Therefore, $f(0) \ge f(\infty) = 0$, which is the EPI (8c).

Note that the case of equality in (8c) is easily determined by this approach, since it reduces to the case of equality in the corresponding FII (see Section II-C–4). Namely, equality holds in the EPI (8c) iff all random vectors X_i for which $a_i \neq 0$ are Gaussian with *identical* covariances. It follows that equality holds in the classical form of the EPI (8a) iff all random vectors X_i for which $a_i \neq 0$ are Gaussian with *proportional* covariances.

2) Integral Representations of Differential Entropy: In the above proof, de Bruijn's identity can be rewritten as an integral representation of entropy. To see this, introduce an auxiliary Gaussian random vector X^*, and rewrite de Bruijn identity (41b) in the form[10] $\frac{d}{dt}(h(X^* + \sqrt{t}\,Z) - h(X + \sqrt{t}\,Z)) = -\frac{1}{2}(J(X + \sqrt{t}\,Z) - J(X^* + \sqrt{t}\,Z))$. Since $h(X^* + \sqrt{t}\,Z) - h(X + \sqrt{t}\,Z) \to 0$ as $t \to \infty$, we may integrate from $t = 0$ to $+\infty$ to obtain $h(X) - h(X^*)$ as the integral of $J(X + \sqrt{t}\,Z) - J(X^* + \sqrt{t}\,Z)$. If, for example, Z^* is chosen standard, one obtains the integral representation [46]

$$h(X) - \frac{n}{2}\log(2\pi e) = -\frac{1}{2}\int_0^\infty J(X + \sqrt{t}\,Z) - \frac{n}{1+t}\,dt. \quad (47a)$$

In view of this identity, the EPI (8c) immediately follows from the corresponding FII (17c).

3) Other Paths of Integration: Several variants of the above proof were published, either in differential or integral form. Dembo, Cover, and Thomas [4] and Carlen and Soffer [5] use a path connecting Z to X of the form $\{\sqrt{t}\,X + \sqrt{1-t}\,Z\}_{t\in(0;1)}$. The argument leading to the EPI is the same up to an appropriate change of variable. The corresponding integral representation

$$h(X) = \frac{n}{2}\log(2\pi e) - \frac{1}{2}\int_0^1 J(\sqrt{t}\,X + \sqrt{1-t}\,Z) - n\,\frac{dt}{t} \quad (47b)$$

was first used by Barron [30] to prove a strong version of the central limit theorem. Verdú and Guo [11] used the path $\{\sqrt{t}\,X + Z\}_{t\in(0;+\infty[}$ and replaced Fisher information by MMSE. They used (42b) to integrate inequality (33) over this path. Their proof is completely equivalent to Stam's proof above, by means of the complementary relation (29) of Proposition 6 and the change of variable $t' = 1/t$. The corresponding integral representation becomes [10]–[12]

$$h(X) = \frac{n}{2}\log(2\pi e) - \frac{1}{2}\int_0^\infty \frac{n}{1+t} - \mathsf{Var}(X \mid \sqrt{t}\,X + Z)\,dt. \quad (47c)$$

[10]When X^* is chosen such that $\mathsf{Cov}(X^*) = \mathsf{Cov}(X)$, the identity relates nonnegative "non-Gaussiannesses" (6) and (15).

Yet another possibility is to take the path $\{\sqrt{1-t}\,X + \sqrt{t}\,Z\}_{t\in(0;1)}$ connecting X to Z, leading to the following integral representation:

$$h(X) = \frac{n}{2}\log(2\pi e) - \frac{1}{2}\int_0^\infty n - \frac{1}{t}\mathsf{Var}(X \mid \sqrt{1-t}\,X + \sqrt{t}\,Z)\,\frac{dt}{t}. \quad (47d)$$

All the above representations for entropy are equivalent through appropriate changes of variable inside the integrals.

III. A New Proof of Shannon's EPI

A. A Mutual Information Inequality (MII)

From the analysis made in Section II, it is clear that earlier information theoretic proofs of the EPI can be seen as variants of the same proof, with the following common ingredients:

1) a *data processing inequality* applied to the linear transformation (9).
2) an integration over a path of a continuous *Gaussian perturbation*.

While step 1) uses the data processing theorem in terms of either parametric Fisher information or MMSE, step 2) uses de Bruijn's identity, which relates Fisher information or MMSE to entropy or mutual information. This suggests that it should be possible to prove the EPI via a data processing argument made directly on the mutual information. The interest is twofold: First, compared to the data processing theorem for Fisher information, the corresponding theorem for Shannon's mutual information is presumably more familiar to the readers of this journal. Second, this approach sidesteps both Fisher information and MMSE and avoids the use of de Bruijn's identity (41b) or (42b).

We shall prove a stronger statement than the EPI, namely, that the difference between both sides of (8c) decreases as independent Gaussian noise Z is added. Since $h(X + Z) - h(X) = I(X + Z; Z)$ for any X independent of Z (see Lemma 1), we write this statement in terms of mutual information as follows.

Theorem 1 (Mutual Information Inequality (MII)): For finitely many independent random n-vectors $(X_i)_i$ with finite covariances, any real-valued coefficients $(a_i)_i$ normalized such that $\sum_i a_i^2 = 1$, and any Gaussian n-vector Z independent of $(X_i)_i$

$$I\Big(\sum_i a_i X_i + Z; Z\Big) \le \sum_i a_i^2 I(X_i + Z; Z). \quad (48)$$

Furthermore, this inequality implies the EPI (8c).

The MII (48) can be interpreted as a convexity property of mutual information under the covariance-preserving transformation (9). As we shall see, the crucial step in the proof of Theorem 1 is the data processing inequality for mutual information [9]. We also need the following technical lemmas. In order to be mathematically correct throughout we first establish some basic properties of mutual information and entropy.

Lemma 1: Let X be any random n-vector and Z be any Gaussian n-vector independent of X. Then $X + Z$ has a density, $h(X + Z)$ exists and is finite. In addition, if $h(X)$ exists, the identity $I(X + Z; Z) = h(X + Z) - h(X)$ always holds.

Proof: Let $\phi_X(u) = \mathsf{E}(\exp(ju \cdot X))$ be the characteristic function of X; that of $Y = X + Z$ is $\phi_Y(u) = \phi_X(u)\phi_Z(u)$ where $\phi_Z(u) = \exp(ju \cdot \mathsf{E}(Z) - \frac{1}{2}u^t \mathsf{Cov}(Z)u)$. Since characteristic functions are bounded continuous and $\phi_Z(u)$ has rapid decay (faster than any inverse of a polynomial) at infinity, it follows that $\phi_Y(u)$ is integrable. Therefore, Y admits a bounded density[11] $p(y)$, such that $p(y) \le c$ where c is some positive constant. The negative part of the integral $-\int p(y)\log p(y)dy$ is $h^-(Y) = \int_{p(y)\ge 1} p(y)\log p(y)dy \le \log c$, which is bounded. Hence $h(X+Z) = h(Y)$ exists and is finite.

If $h(X)$ exists, then either X admits a density $p_X(x)$ or it does not. In the former case $(Y,Z) = (X+Z,Z)$ also admits a density $p(y,z) = p_X(y-z)p_Z(z)$ and the identity $I(X+Z;Z) = h(X+Z) - h(X)$ is well known. In the latter case we have put $h(X) = -\infty$ (see Section I-B). Since X is not absolutely continuous with respect to the Lebesgue measure, there exists a set A of zero measure such that $P(X \in A) > 0$. Then $B = \{(y,z) \,|\, y - z \in A\}$ has zero Lebesgue measure and $P((X+Z,Z) \in B) = P(X \in A) > 0$. Since B is also of zero measure with respect to the product probability measure with density $p_Y(y)p_Z(z)$, it follows that (Y,Z) is not absolutely continuous with respect to this product measure. Therefore, by the theorem of Gel'fand-Yaglom-Perez [50, ch. 2], one has $I(X+Z;Z) = I(Y;Z) = +\infty$ and the identity $I(X+Z;Z) = h(X+Z) - h(X)$ still holds. □

In the same way one can prove that the identity $I(X + \sqrt{t}\,Z;Z) = h(X+\sqrt{t}\,Z) - h(X)$ always holds for any $t \ge 0$.

The following inequality (49) was proved for two variables by Sato [86] who used it to derive an outer bound to the capacity region of broadcast channels. A similar inequality appears in [87, Th. 4.2.1] and in [88, Th 1.9].

Lemma 2 (Sato's Inequality): If the random vectors $(X_i)_i$ are independent of Z and of each other, then

$$I((X_i+Z)_i;Z) \le \sum_i I(X_i+Z;Z). \tag{49}$$

Proof: Let $Y_i = X_i + Z$ for all i. By the chain rule for mutual information [50, ch. 3], one has

$$\begin{aligned} I((Y_i)_i;Z) &= \sum_i I(Y_i;Z \,|\, Y_1,\dots,Y_{i-1}) && (50a)\\ &\le \sum_i I(Y_i;Z \,|\, Y_1,\dots,Y_{i-1}) + I(Y_i;Y_1,\dots,Y_{i-1}) && (50b)\\ &= \sum_i I(Y_i;Z,Y_1,\dots,Y_{i-1}) && (50c)\\ &= \sum_i I(Y_i;Z) + I(Y_i;Y_1,\dots,Y_{i-1} \,|\, Z) && (50d)\\ &= \sum_i I(Y_i;Z). && (50e) \end{aligned}$$

□

[11]This density is in fact indefinitely differentiable and strictly positive, and all its derivatives are bounded and tend to zero at infinity.

An alternative proof in the case where $\mathbf{Y} = (Y_i)_i$ admits a density, is as follows. Define the *multi-information* between the components of $\mathbf{Y}$ by the divergence

$$I\{(Y_i)_i\} = \mathsf{E}\log\frac{p(\mathbf{Y})}{\prod_i p(Y_i)} = \sum_{i>1} I(Y_i;Y_1,\dots,Y_{i-1}). \tag{51}$$

From the definitions it is obvious that $I((Y_i)_i;Z) - \sum_i I(Y_i;Z) = I\{(Y_i)_i \,|\, Z\} - I\{(Y_i)_i\}$. The result follows since $I\{(Y_i)_i\} \ge 0$ and $I\{(Y_i)_i \,|\, Z\} = I\{(X_i)_i\} = 0$.

Lemma 3: If X and Z are independent random n-vectors with finite covariances and differential entropies, then

$$\lim_{t\to 0^+} I(X + \sqrt{t}Z;Z) = 0. \tag{52}$$

If, in addition, $I(X+\sqrt{t}Z;Z)$ is differentiable at $t=0$, then for any real constant a

$$I(X + a\sqrt{t}Z;Z) = a^2 I(X+\sqrt{t}Z;Z) + o(t) \tag{53}$$

where $o(t)$ is a function defined for all $t \ge 0$ such that $o(t)/t \to 0$ as $t \to 0^+$.

Proof: To prove (52), let $X_t = X + \sqrt{t}\,Z$. Taking characteristic functions, $\phi_{X_t}(u) = \phi_X(u)\phi_Z(\sqrt{t}u) \to \phi_X(u)$ as $t \to 0^+$. Therefore $X_t \to X$ in distribution. Let X^* and Z^* be Gaussian n-vectors have identical covariances as X and Z, respectively. Likewise $X_t^* = X^* + \sqrt{t}Z^* \to X^*$ in distribution. By the lower semicontinuity of divergence (see [50, Sec. 2.4] and [89, Th. 1]), we have

$$D(X\|X^*) \le \liminf_{t\to 0^+} D(X_t\|X_t^*). \tag{54}$$

The following quantities are all finite.

$$\begin{aligned} &D(X_t\|X_t^*) - D(X\|X^*)\\ &\quad = h(X_t^*) - h(X_t) - h(X^*) + h(X^*) && (55a)\\ &\quad = h(X^* + \sqrt{t}Z^*) - h(X^*) - I(X+\sqrt{t}Z;Z). && (55b) \end{aligned}$$

An easy calculation for Gaussian vectors gives $\lim_{t\to 0^+} h(X^* + \sqrt{t}Z^*) = h(X^*)$. Therefore (54) reduces to

$$\limsup_{t\to 0^+} I(X+\sqrt{t}Z;Z) \le \liminf_{t\to 0^+} h(X^*+\sqrt{t}Z^*) - h(X^*) = 0. \tag{56}$$

This combined with nonnegativity of mutual information proves (52).

Now suppose $I(X+\sqrt{t}Z;Z)$ is differentiable at $t=0$. Since $\lim_{t\to 0^+} I(X+\sqrt{t}\,Z;Z) = 0$, for any $a \in \mathbb{R}$, $I(X + a\sqrt{t}Z;Z)/a^2 t$ and $I(X+\sqrt{t}Z;Z)/t$ tend toward the same limit as $t \to 0^+$. This reduces to (53). □

Note that neither Lemma 2 nor Lemma 3 requires Z to be Gaussian. The following lemma gives an important situation where the differentiability assumption of Lemma 3 is met.

Lemma 4: Let X be any random n-vector with finite covariances and differential entropy, and let Z, Z' be identically distributed Gaussian n-vectors such that X, Z, Z' are independent. The quantity $I(X+\sqrt{t}Z;Z)$ is differentiable at any $t>0$. In addition, if $X' = X + \sqrt{u}\,Z'$ where $u>0$, then

$$I(X'+\sqrt{t}Z;Z) = I(X+\sqrt{u+t}\,Z;Z) - I(X+\sqrt{u}\,Z;Z) \quad (57)$$

is also differentiable at $t=0$.

Proof: Following Stam [2], Barron [30] proved that $h(X+\sqrt{t}Z)$ is differentiable in $t>0$ for any square-integrable X. The proof involves exchanges of differentiation and expectation justified by the dominated convergence theorem and is not repeated here. From Lemma 1 is follows that $I(X+\sqrt{t}Z;Z)$ is likewise differentiable at any $t>0$. Now the following quantities are all finite.

$$\begin{aligned}
&I(X'+\sqrt{t}Z;Z)\\
&\quad= I(X+\sqrt{u}\,Z'+\sqrt{t}\,Z;Z) && (58a)\\
&\quad= h(X+\sqrt{u}\,Z'+\sqrt{t}\,Z) - h(X+\sqrt{u}\,Z') && (58b)\\
&\quad= h(X+\sqrt{u}\,Z'+\sqrt{t}\,Z) - h(X)\\
&\qquad + h(X) - h(X+\sqrt{u}\,Z') && (58c)\\
&\quad= I(X+\sqrt{u}\,Z'+\sqrt{t}\,Z;\sqrt{u}\,Z'+\sqrt{t}\,Z)\\
&\qquad - I(X+\sqrt{u}\,Z';\sqrt{u}\,Z') && (58d)\\
&\quad= I(X+\sqrt{u+t}\,Z;Z) - I(X+\sqrt{u}\,Z;Z). && (58e)
\end{aligned}$$

The last equality follows from the stability property of the Gaussian distribution under convolution, since $\sqrt{u}\,Z' + \sqrt{t}\,Z$ is identically distributed as $\sqrt{u+t}\,Z$. Since $I(X+\sqrt{u+t}\,Z;Z)$ is differentiable at $t=0$ for any $u>0$, $I(X'+\sqrt{t}\,Z;Z) = I(X+\sqrt{u+t}\,Z;Z) - I(X+\sqrt{u}\,Z;Z)$ is likewise differentiable at $t=0$. □

Proof of Theorem 1: We may always assume that $a_i \neq 0$ for all i—otherwise simply delete the X_i for which $a_i = 0$. To prove (48), we may also assume that all the X_i have finite differential entropies, since otherwise the right-hand side (RHS) of (48) is $= +\infty$ by Lemma 1. Then all the X_i admit densities, and $\sum_i a_i X_i$ likewise admits a density and has finite covariances. From Proposition 1 it follows that $h(\sum_i a_i X_i) < +\infty$, and since conditioning reduces entropy, $-\infty < h(X_i) \leq h(\sum_i a_i X_i)$. Therefore, $\sum_i a_i X_i$ also has finite differential entropy. From this and Lemma 1 it follows that all subsequent mutual informations will be finite.

We can write the following string of inequalities:

$$\begin{aligned}
I\big(\sum_i a_i X_i + Z;Z\big) &= I\big(\sum_i a_i(X_i + a_i Z);Z\big) && (59a)\\
&\leq I((X_i + a_i Z)_i;Z) && (59b)\\
&\leq \sum_i I(X_i + a_i Z;Z) && (59c)
\end{aligned}$$

where (59a) holds since $\sum_i a_i^2 = 1$, (59b) follows from the data processing theorem applied to the linear transformation (9), (59c) follows from Sato's inequality (Lemma 2). Note that substituting $\sqrt{t}Z$ for Z in (59c) and assuming that Z and the X_i satisfy the differentiability assumption of Lemma 3 for all i, one obtains

$$I\big(\sum_i a_i X_i + \sqrt{t}Z;Z\big) \leq \sum_i a_i^2 I(X_i + \sqrt{t}Z;Z) + o(t). \quad (60)$$

We now use the assumption that Z is *Gaussian* to eliminate the $o(t)$ term in (60).

Let $X_i' = X_i + \sqrt{u}\,Z_i'$ for all i and $u>0$, where the Z_i' are Gaussian, identically distributed as Z but independent of all other random vectors. Then $Z' = \sum_i a_i Z_i'$ is identically distributed as Z, and applying (59c) to the X_i' and to $\sqrt{t}\,Z$, one obtains

$$\begin{aligned}
&I\big(\sum_i a_i X_i + \sqrt{u}\,Z' + \sqrt{t}Z;Z\big)\\
&\quad\leq \sum_i I(X_i' + a_i Z;Z) && (61a)\\
&\quad= \sum_i a_i^2 I(X_i + \sqrt{u}\,Z_i' + \sqrt{t}Z;Z) + o(t) && (61b)
\end{aligned}$$

where the last equality follows from the fact that by Lemma 4, the $X_i' = X_i + \sqrt{u}\,Z_i'$ satisfy the differentiability assumption of Lemma 3. Now define

$$f(t) = I\big(\sum_i a_i X_i + \sqrt{t}Z;Z\big) - \sum_i a_i^2 I(X_i + \sqrt{t}Z;Z).$$

Using (57), inequality (61) is easily rewritten as

$$\begin{aligned}
&I\big(\sum_i a_i X_i + \sqrt{u+t}\,Z;Z\big) - I\big(\sum_i a_i X_i + \sqrt{u}\,Z;Z\big)\\
&\leq \sum_i a_i^2 (I(X_i + \sqrt{u+t}\,Z;Z) - I(X_i + \sqrt{u}\,Z;Z)) + o(\varepsilon)
\end{aligned}$$

that is $f(u+t) \leq f(u) + o(t)$ for any $u>0$. Since $f(u)$ is differentiable at any $u>0$ by Lemma 4, it easily follows that $f(u)$ is non-increasing in $u>0$. Also, by Lemma 3, $\lim_{t\to 0} f(t) = f(0) = 0$. Therefore, $f(1) \leq f(0) = 0$, which is the required MII (48).

Finally, we show that the MII implies the EPI (8c). Since $\sum_i a_i^2 = 1$ and $I(X+Z;Z) = h(X+Z) - h(X) = I(X;X+Z) + h(Z) - h(X)$ for X independent of Z, (48) can be rewritten as

$$\begin{aligned}
h\big(\sum_i a_i X_i\big) - \sum_i a_i^2 h(X_i) &\geq I\big(\sum_i a_i X_i;\sum_i a_i X_i + Z\big)\\
&\quad - \sum_i a_i^2 I(X_i;X_i+Z). && (62)
\end{aligned}$$

Now replace Z by $\sqrt{t}\,Z$ and let $t\to\infty$. The terms in the RHS of the above inequality are of the form $I(X;X+\sqrt{t}\,Z) = I(X;\frac{1}{\sqrt{t}}X + Z)$, which tends to zero as $t\to\infty$ by Lemma 3. This completes the proof. □

B. Insights and Discussions

1) Relationship to Earlier Proofs: Of course, Theorem 1 could also be proved using the conventional techniques of Section II. In fact, it follows easily from either one of the

integral representations (47). Also Lemma 3 is an easy consequence of de Bruijn's identity, since by (37), both sides of (53) are equal to $\frac{t}{2}\mathsf{tr}(\mathsf{Cov}(aZ)\mathbf{J}(X)) = \frac{a^2t}{2}\mathsf{tr}(\mathsf{Cov}(Z)\mathbf{J}(X))$. The originality here lies in the above proof of Theorem 1 and the EPI, which in contrast to existing proofs, requires neither de Bruijn's identity nor the notions of Fisher information or MMSE.

The new proof shares common ingredients with earlier proofs of the EPI, namely items 1) and 2) listed at the beginning of this section. The difference is that they are used directly in terms of mutual information. As in Section II-E–3, other paths of continuous Gaussian perturbation could very well be used, through suitable changes of variable.

One may wonder if mutual informations in the form $I(X;\sqrt{t}\,X+Z)$ rather than $I(X+\sqrt{t}\,Z;Z)$ could be used in the above derivation of Theorem 1, particularly in inequalities (59). This would offer a dual proof, in the same way as Verdú and Guo's proof is dual to Stam and Blachman's original proof of the EPI, as explained in Section II. But a closer look at the above proof reveals that the dual approach would amount to prove (62), whose natural proof using the data processing inequality is through (59). Thus, it turns out that the two approaches amount to the same.

Also note that by application of de Bruijn's identity, inequality (60) reduces to the FII (17c). Thus the MII (48) implies both the EPI (8c) and the FII (17c).

2) The Equality Case: Our method does not easily settle the case of equality in the MII. By the preceding remark, however, equality in (60) implies equality in the FII (17c), which was determined in Section II-C–4. It follows that equality holds in the MII (48) if and only if all random vectors X_i such that $a_i \neq 0$ are Gaussian with identical covariances. This result implies the corresponding necessity condition of equality in the EPI, but is not evident from the properties of mutual information alone.

3) On the Gaussianness of Z: It is interesting to note that from (60), the MII holds up to first order of the noise variance, regardless of whether Z is Gaussian or not. However, the stability property of the Gaussian distribution under convolution was crucial in the next step of the proof, because the Gaussian perturbation Z can be made to affect the random vectors independently. In fact, the MII can be easily rewritten as

$$h\Big(\sum_i a_i X_i\Big) - \sum_i a_i^2 h(X_i) \geq h\Big(\sum_i a_i X_i'\Big) - \sum_i a_i^2 h(X_i') \quad (63)$$

where $X_i' = X_i + Z_i$ for all i, the $(Z_i)_i$ being independent copies of Z. This does *not* hold in general for non-Gaussian random vectors $(Z_i)_i$. To see this, choose $(X_i)_i$ themselves Gaussian with identical covariances. Then the left-hand side (LHS) of (63) is zero, and by the necessity of the condition for equality in the EPI, the RHS is positive, as soon as Z_i is non-Gaussian for some i such that $a_i \neq 0$. Therefore, in this case, the opposite inequality is obtained. In other words, adding non-Gaussian noise may *increase* the difference between both sides of the EPI (8c), in accordance with the fact that this difference is zero for Gaussian random vectors.

4) On the Finite Second-Order Moment Assumption: To prove Theorem 1 we have assumed for simplicity that the X_i have finite covariances so that differential entropies are well defined and the lower semicontinuity argument in Lemma 3 applies. However, it would be possible to weaken this condition to first-order finite moment or even to the condition of Proposition 1 by considering divergences with respect to probability distributions other than Gaussian, e.g., exponential or Cauchy distributions as in the proof of Proposition 1. The details are left to the reader.

5) On the Use of Sato's Inequality: Sato used (49) and the data processing inequality to derive his cooperative outer bound to the capacity region of two-user broadcast channels [86]. This bound was used to determine the capacity of a two-user MIMO Gaussian broadcast channel [90]. Sato's bound was later replaced by the EPI to generalize Bergmans' solution to an arbitrary multiuser MIMO Gaussian broadcast channel using the notion of an "enhanced" channel [15]. In the present paper, the EPI itself is proved using Sato's inequality and the data processing inequality. This suggests that for proving converse coding theorems, a direct use of the EPI may be avoided by suitable inequalities for mutual information. A similar remark goes for the generalization of Ozarow's solution to vector Gaussian multiple descriptions [91].

6) Relationship Between Various Data Processing Theorems: Proposition 5 enlightens the connection between two estimation theoretic data processing inequalities: parametric (Fisher information) and nonparametric (MMSE). While these were applied in earlier proofs of the EPI, the new proof uses the same data processing argument in terms of mutual information: any transformation $X \to Y$ in a Markov chain $\theta \to X \to Y$ reduces information about θ. This can also be given a parametric form using divergence (38). Thus, if $\theta \to X \to Y$ form a Markov chain, then

$$I(\theta,Y) \leq I(\theta,X) \quad (64a)$$

$$D_Y(p_\theta \| p_{\theta'}) \leq D_X(p_\theta \| p_{\theta'}). \quad (64b)$$

As in Proposition 5, the first data processing inequality involves a random variable θ, while the second considers θ as a parameter. The proof is immediate from the chain rules $I(\theta;Y) + I(\theta;X\,|\,Y) = I(\theta;X,Y) = I(\theta;X)$ and $D_Y(p_\theta\|p_{\theta'}) + D_{X\,|\,Y}(p_\theta\|p_{\theta'}) = D_{X,Y}(p_\theta\|p_{\theta'}) = D_X(p_\theta\|p_{\theta'})$ where by the Markov chain condition, $I(\theta;Y\,|\,X)=0$ and $D_{Y\,|\,X}(p_\theta\|p_{\theta'})=0$, respectively.

Comparing the various proofs of the EPI presented above, it is clear that, as already suggested in Zamir's presentation [53], estimation theoretic and information theoretic data processing inequalities are strongly related. Also note that in view of (39), the lesser known data processing inequality for Fisher information (23b) is an immediate consequence of the corresponding inequality for divergence (64b). Indeed, dividing both sides of (64b) by $\|\theta-\theta'\|^2$ and letting $\theta' \to \theta$ gives (23b). It would be interesting to see if the various data processing inequalities (for mutual information, divergence, MMSE, and Fisher information) can be further unified and given a common viewpoint, leading to new insights and applications.

7) On the EPI for Discrete Variables: The above proof of the MII does not require the $(X_i)_i$ to be random vectors with densities. Therefore, it also holds when the random vectors are discrete (finitely or countably) valued. In fact, Verdú and Guo [11] used [10, Lemma 6, App. VII] to show that the EPI (8c)

also holds in this case, where differential entropies are replaced by entropies. We call attention that this is in fact a immediate consequence of the stronger inequality

$$H\big(\sum_i a_i X_i\big) \geq \max_i H(X_i)$$

for any independent discrete random vectors $(X_i)_i$ and any real-valued coefficients $(a_i)_i$, which is easily obtained by noting that $H(\sum_i a_i X_i) \geq H(\sum_i a_i X_i \,|\, (X_j)_{j\neq i}) = H(X_i)$ for all i. Note, however, that the classical EPI in the form $\exp\frac{2}{n}H(\sum_i X_i) \geq \sum_i \exp\frac{2}{n}H(X_i)$ does *not* hold in general for discrete random vectors—a simple counterexample is obtained by taking deterministic X_i for all i.

There also exist may discrete analogs to the entropy power inequality, either in the form (8a) or (8c). A first set of results [92]–[95] were derived for binary random vectors where addition is replaced by modulo-2 addition. The corresponding inequalities are quite different from (8) and apparently unrelated to the contributions of this paper.

More recent results involve random variables taking integer values. In this case, the role of the Gaussian distribution and it stability property under convolution is played by the Poisson distribution. An analog of the FII (17) was proposed by Kagan [96] and a similar, alhtough different, version of discrete Fisher information was used in [97] in connection with the convergence of the (usual) sum of independent binary random variables toward the Poisson distribution. A discrete analog to (8a) was proved for binomial distributions [98], and a discrete analog to (8c) was recently established by Yu and Johnson [99] for positive random variables having ultra-log-concave distributions. It would be desirable to unify the different approaches for integer-valued random variables to see whether the method of this paper contributes to what is known in this case.

IV. Zamir and Feder's EPI For Linear Transformations

A. Background

Zamir and Feder [38]–[40] generalized the scalar EPI by extending the linear combination $\sum_i a_i X_i$ of random variables to an arbitrary linear transformation $\mathbf{A}X$, where X is the random vector of independent entries $(X_j)_j$ and $\mathbf{A} = (a_{i,j})_{i,j}$ is a rectangular matrix. They showed that the resulting inequality cannot be derived by a straightforward application of the vector EPI of Proposition 2. They also noted that it becomes trivial if $\mathbf{A}$ is row-rank deficient. Therefore, in the following, we assume that $\mathbf{A}$ has full row rank.

Zamir and Feder's generalized EPI (ZF-EPI) has been used to derive results on closeness to normality after linear transformation of a white random vector in the context of minimum entropy deconvolution [38] and analyze the rate-distortion performance of an entropy-coded dithered quantization scheme [100]. It was also used as a guide to extend the Brunn-Minkowski inequality in geometry [101], [102], which can be applied to the calculation of lattice quantization bit rates under spectral constraints.

The equivalent forms of the ZF-EPI corresponding to those given in Proposition 2 are the following.

Proposition 9 (Equivalent ZF-EPIs): The following inequalities, each stated for any random (column) vector X of independent entries $(X_j)_j$ with densities and real-valued rectangular full row rank matrix $\mathbf{A}$, are equivalent.

$$N(\mathbf{A}X) \geq |\mathbf{A}\,\mathsf{diag}(N(X_j))_j\,\mathbf{A}|^{1/r} \tag{65a}$$

$$h(\mathbf{A}X) \geq h(\mathbf{A}\tilde{X}) \tag{65b}$$

$$h(\mathbf{A}X) \geq \sum_{i,j} a_{i,j}^2 h(X_j) \quad (\mathbf{A}\mathbf{A}^t = \mathbf{I}) \tag{65c}$$

where r is the number of rows in $\mathbf{A}$, and the components of $\tilde{X} = (\tilde{X}_j)_j$ are independent Gaussian random variables of entropies $h(\tilde{X}_j) = h(X_j)$.

The proof of Proposition 9 is a direct extension of that of Proposition 2. That (65a), (65b) are equivalent follows immediately from the equalities $|\mathbf{A}\,\mathsf{diag}(N(X_j))_j\,\mathbf{A}|^{1/r} = |\mathbf{A}\,\mathsf{diag}(N(\tilde{X}_j))_j\,\mathbf{A}|^{1/r} = |\mathbf{A}\mathsf{Cov}(\tilde{X})\mathbf{A}^t|^{1/r} = |\mathsf{Cov}(\mathbf{A}\tilde{X})|^{1/r} = N(\mathbf{A}\tilde{X})$. The implication (65c) $\Longrightarrow$ (65a) is proved in [40], and the equivalence (65b) $\Longleftrightarrow$ (65c) is proved in detail in [12].

Similarly as for (8c), inequality (65c) can be interpreted as a concavity property of entropy under the variance-preserving[12] transformation

$$X \to \mathbf{A}X \quad (\mathbf{A}\mathbf{A}^t = \mathbf{I}) \tag{66}$$

and is the golden door in the route of proving the ZF-EPI. The conventional techniques presented in Section II generalize to the present situation. One has the following Fisher information matrix inequalities analogous to (65):

$$\mathbf{J}^{-1}(\mathbf{A}X) \geq \mathbf{A}\mathbf{J}^{-1}(X)\mathbf{A}^t \tag{67a}$$

$$\mathbf{J}(\mathbf{A}X) \leq \mathbf{J}(\mathbf{A}\hat{X}) \tag{67b}$$

$$\mathbf{J}(\mathbf{A}X) \leq \mathbf{A}\mathbf{J}(X)\mathbf{A}^t \quad (\mathbf{A}\mathbf{A}^t = \mathbf{I}) \tag{67c}$$

where the components of $\hat{X} = (\hat{X}_j)_j$ are independent Gaussian variables with Fisher informations $J(\hat{X}_j) = J(X_j)$ for all j. The first inequality (67a) was derived by Papathanasiou [52] and independently by Zamir and Feder [38], [40], who used a generalization of the conditional mean representation of score (see Section II-C–2); their proof is simplified in [103] and [104]. Later, Zamir [53] provided an insightful proof of (67) by generalizing Stam's approach (see Section II-C–1) and also determined the case of equality [103], [105]. Taking the trace in both sides of (67c) gives

$$J(\mathbf{A}X) \leq \sum_{i,j} a_{i,j}^2 J(X_j) \qquad (\mathbf{A}\mathbf{A}^t = \mathbf{I}) \tag{68}$$

which was used by Zamir and Feder [40], [53] to prove the ZF-EPI (65c) by integration over the path $\{\sqrt{t}X + \sqrt{1-t}Z\}$ (see Section II-E). Finally, Guo, Shamai, and Verdu [12] generalized their approach (see Section II-C–3) to obtain

$$\mathsf{Var}(\mathbf{A}X \,|\, \mathbf{A}X + Z) \geq \sum_{i,j} a_{i,j}^2 \mathsf{Var}(X_j \,|\, X_j + Z_j) \tag{69}$$

where Z and the $(Z_j)_j$ are standard Gaussian independent of X, and used it to prove the ZF-EPI (65c) by integration over the path $\{\sqrt{t}X + Z\}$ (see Section II-E). Again the approaches

[12]If the $(X_j)_j$ have equal variances, then so have the components of $\mathbf{A}X$, since $\mathsf{Cov}(X) = \sigma^2\mathbf{I}$ implies $\mathsf{Cov}(\mathbf{A}X) = \mathbf{A}\mathsf{Cov}(X)\mathbf{A}^t = \sigma^2\mathbf{A}\mathbf{A}^t = \sigma^2\mathbf{I}$.

corresponding to (68) and (69) are equivalent by virtue of the complementary relation (29), as explained in Section II-C–3.

B. A New Proof of the ZF-EPI

The same ideas as in the proof of Theorem 1 are easily generalized to prove the ZF-EPI.

Theorem 2 (Mutual Information Inequality for Linear Transformations): For any random vector X with independent entries $(X_j)_j$ having finite variances, any real-valued rectangular matrix $\mathbf{A}$ with r orthonormal rows ($\mathbf{A}\mathbf{A}^t = \mathbf{I}$), and any standard Gaussian random r-vector $\mathbf{Z}$ and variable Z independent of X

$$I(\mathbf{A}X + \mathbf{Z}; \mathbf{Z}) \le \sum_{i,j} a_{i,j}^2 I(X_j + Z; Z). \tag{70}$$

Furthermore, this inequality imply the ZF-EPI.

Proof: Noting $Z' = \mathbf{A}^t\mathbf{Z}$, a Gaussian random vector with the same dimension as X, we can write the following string of inequalities:

$$\begin{aligned} I(\mathbf{A}X + \mathbf{Z}; \mathbf{Z}) &= I(\mathbf{A}(X + Z'); Z') && \text{(71a)}\\ &\le I(X + Z'; Z') && \text{(71b)}\\ &\le \sum_j I(X_j + Z_j'; Z_j') && \text{(71c)} \end{aligned}$$

where (71a) holds since $\mathbf{A}\mathbf{A}^t = \mathbf{I}$, (71b) follows from the data processing theorem applied to the linear transformation (66), and (71c) follows from Sato's inequality (Lemma 2). Now apply the resulting inequality to $\hat{X} = X + \sqrt{u}\,\hat{Z}$, where $u > 0$ and $\hat{Z}$ is a standard Gaussian random vector independent of all other random variables, and replace $\mathbf{Z}$ by $\sqrt{t}\,\mathbf{Z}$, where $t > 0$. This gives

$$\begin{aligned} &I(\mathbf{A}X + \sqrt{u}\,\mathbf{A}\hat{Z} + \sqrt{t}\,\mathbf{Z}; \mathbf{Z})\\ &\qquad \le \sum_j I(X_j + \sqrt{u}\,\hat{Z}_j + \sqrt{t}\,Z_j'; Z_j'). \end{aligned}$$

The Gaussian perturbation $\hat{Z}$ ensures that densities of the $(X_j + \sqrt{u}\,\hat{Z}_j)_j$ are smooth, so that (52) of Lemma 3 applies to the RHS. Noting that $\mathsf{Cov}(Z') = \mathbf{A}^t\mathbf{A}$ and therefore, $\sigma_{Z_j'}^2 = \sum_i a_{i,j}^2$ for all j, we obtain

$$\begin{aligned} &I(\mathbf{A}X + \sqrt{u}\,\mathbf{A}\hat{Z} + \sqrt{t}\,\mathbf{Z}; \mathbf{Z})\\ &\qquad \le \sum_{i,j} a_{i,j}^2 I(X_j + \sqrt{u}\,\hat{Z}_j + \sqrt{t}\,Z; Z) + o(t) \end{aligned}$$

where $\mathbf{A}\hat{Z}$ is identically distributed as $\mathbf{Z}$ (since $\mathbf{A}\mathbf{A}^t = \mathbf{I}$), and Z is a standard Gaussian variable, independent of all other random variables. By the stability property of the Gaussian distribution under convolution, $\sqrt{u}\,\mathbf{A}\hat{Z} + \sqrt{t}\,\mathbf{Z}$ is identically distributed as $\sqrt{u+t}\,\mathbf{Z}$, and the $(\sqrt{u}\,\hat{Z}_j + \sqrt{t}\,Z)_j$ are identically distributed as $\sqrt{u+t}\,Z$. Therefore, applying (58) gives

$$\begin{aligned} &I(\mathbf{A}X + \sqrt{u+t}\,\mathbf{Z}; \mathbf{Z}) - I(\mathbf{A}X + \sqrt{u}\,\mathbf{Z}; \mathbf{Z})\\ &\le \sum_{i,j} a_{i,j}^2 (I(X_j + \sqrt{u+t}Z; Z) - I(X_j + \sqrt{u+t}Z; Z)) + o(t) \end{aligned}$$

which shows that

$$f(t) = I(\mathbf{A}X + \sqrt{t}\,\mathbf{Z}; \mathbf{Z}) - \sum_{i,j} a_{i,j}^2 I(X_j + \sqrt{t}\,Z; Z)$$

is nonincreasing in $t > 0$. Also, by Lemma 3, $\lim_{t\to 0} f(t) = f(0) = 0$. Therefore, $f(1) \le f(0) = 0$, which proves the required MII (70).

Finally, we show that (70) implies the ZF-EPI (65c). By the identity $I(X+Z; Z) = I(X; X+Z) + h(Z) - h(X)$ for any X independent of Z, the MII in the form $f(t) \le 0$ can be rewritten as

$$\begin{aligned} &h(\mathbf{A}X) - \sum_{i,j} a_{i,j}^2 h(X_j)\\ &\qquad \ge I(\mathbf{A}X; \mathbf{A}X + \sqrt{t}\,\mathbf{Z})\\ &\qquad\quad - \sum_{i,j} a_{i,j}^2 I(X_j; X_j + \sqrt{t}\,Z) + \Delta \end{aligned}$$

where $\Delta = h(\sqrt{t}\,\mathbf{Z}) - \sum_{i,j} a_{i,j}^2 h(\sqrt{t}\,Z) = rh(\sqrt{t}\,Z) - rh(\sqrt{t}\,Z) = 0$. The other terms in the RHS of this inequality are of the form $I(X + \sqrt{t}\,Z; Z) = I(X; \frac{1}{\sqrt{t}}X + Z)$, which letting $t \to \infty$ tends to zero by Lemma 3. This completes the proof. □

Notice that the approach presented here for proving the ZF-EPI is the same as for proving the original EPI, namely, that the difference between both sides of the ZF-EPI (65c) is decreased as independent white Gaussian noise is added

$$h(\mathbf{A}X) - \sum_{i,j} a_{i,j}^2 h(X_j) \ge h(\mathbf{A}X') - \sum_{i,j} a_{i,j}^2 h(X_j') \tag{72}$$

where $X' = X + Z$ and Z is white Gaussian independent of X.

Zamir and Feder derived their results for random variables X_j. However, our approach can be readily extended to random n-vectors. For this purpose, consider the random vector $X = (X_j)_j$ whose components X_j are themselves n-vectors, and adopt the convention that the components of $Y = \mathbf{A}X$ are n-vectors given by the relations $Y_i = \sum_j a_{i,j}X_j$, which amounts to saying that $\mathbf{A}$ is a block matrix with submatrix entries $(a_{i,j}\mathbf{I})_{i,j}$. The generalization of Theorem 2 is straightforward and we omit the details. The corresponding general ZF-EPI is still given by (65), with the above convention in the notations.

V. Takano and Johnson's EPI for Dependent Variables

A. Background

Takano [41] and Johnson [42] provided conditions under which the EPI, in the form $N(X_1 + X_2) \ge N(X_1) + N(X_2)$, would still hold for dependent variables. These conditions are expressed in terms of appropriately perturbed variables

$$X_{i,t} = X_i + \sqrt{f_i(t)}\,Z_i \quad (i = 1, 2) \tag{73}$$

where Z_1, Z_2 are standard Gaussian, independent of $X = (X_1, X_2)^t$ and of each other, and $f_1(t)$ and $f_2(t)$ are positive functions which tend to infinity as $t \to \infty$. They involve individual scores $S(X_{1,t}), S(X_{2,t})$ and Fisher informations $J(X_{1,t}), J(X_{2,t})$, as well as the entries of the joint score $S(X_t) = (S_1(X_t), S_2(X_t))^t$ and the Fisher information matrix

$$\mathbf{J}(X_t) = \begin{pmatrix} J_{1,1}(X_t) & J_{1,2}(X_t) \\ J_{1,2}(X_t) & J_{2,2}(X_t) \end{pmatrix}$$

where $X_t = (X_{1,t}, X_{2,t})^t$. Takano's condition is [41]

$$2\frac{\mathsf{E}(S(X_{1,t})S(X_{2,t}))}{J(X_{1,t})J(X_{2,t})} \geq \mathsf{E}\left(\left\{\frac{S_1(X_t)-S(X_{1,t})}{J(X_{1,t})}+\frac{S_2(X_t)-S(X_{2,t})}{J(X_{2,t})}\right\}^2\right) \quad (74)$$

for all $t > 0$. Johnson's improvement is given by the following weaker condition [42]:

$$2\frac{\mathsf{E}\left(S(X_{1,t})S(X_{2,t})\right)}{J(X_{1,t})J(X_{2,t})} \geq \mathsf{E}\left(\left\{\frac{(J_{2,2}(X_t)-J_{1,2}(X_t))\,S_1(X_t)+(J_{1,1}(X_t)-J_{1,2}(X_t))\,S_2(X_t)}{J_{1,1}(X_t)J_{2,2}(X_t)-J_{1,2}^2(X_t)} - \frac{S(X_{1,t})}{J(X_{1,t})}-\frac{S(X_{2,t})}{J(X_{2,t})}\right\}^2\right) \quad (75)$$

for all $t > 0$. These conditions were found by generalizing the conventional approach presented in Section II, in particular Blachman's representation of the score (Section II-C–2). They are simplified below. The EPI for dependent variables finds its application in entropy-based blind source separation of dependent components (see, e.g., [106]).

B. A Generalized EPI for Dependent Random Vectors

In this section, we extend Theorem 1 to provide a simple condition on dependent random n-vectors $(X_i)_i$ under which not only the original EPI $N(\sum_i X_i) \geq \sum_i N(X_i)$ holds, but also the EPI (8) for *any* choice of coefficients $(a_i)_i$. Such stronger form should be more relevant in applications such as blind separation of dependent components, for it ensures that negentropy $-h$ still satisfies the requirements (10) for a contrast objective function, for any type of linear mixture. Define

$$X_{i,t} = X_i + \sqrt{t}\,Z_i \quad (76)$$

corresponding to (73) with $f_i(t) = t$ for all i. Our condition will be expressed in terms of *symmetric mutual information* $I\{(X_{i,t})_i\} \geq 0$ defined by (51), which serves as a measure of dependence between the components of a random vector.

Theorem 3: Let $X = (X_i)_i$ be any finite set of (dependent) random n-vectors, let $X_t = (X_{i,t})_i$ be defined by (76), and let Z be a white Gaussian random n-vector independent of all other random vectors. If, for any $t > 0$ and any real-valued coefficients $(a_i)_i$, adding a small perturbation $a_i Z$ to the $X_{i,t}$ makes them "more dependent" in the sense that

$$I\{(X_{i,t}+a_iZ)_i\} \geq I\{(X_{i,t})_i\} + o\left(\sigma_Z^2\right) \quad (77)$$

then the MII (48) and the EPI (8) hold for these random vectors $(X_i)_i$.

Proof: The only place where the independence of the $(X_i)_i$ is used in the proof of Theorem 1 is Sato's inequality (59c), which is used to the first order of σ_Z^2 and applied to random vectors of the form (76) for all $t > 0$. Therefore, it is sufficient that

$$I((X_{i,t}+a_iZ)_i; Z) \leq \sum_i I(X_{i,t}+a_iZ; Z) + o(\sigma_Z^2)$$

holds for all $t > 0$ and any choice of $(a_i)_i$ to prove the MII and hence the EPI. Now from the proof of Lemma 2, the difference between both sides of this inequality is

$$I((X_i+a_iZ)_i; Z) - \sum_i I(X_i+a_iZ; Z) = I\{(X_i)_i\} - I\{(X_i+a_iZ)_i\}.$$

The result follows at once. □

Note that it is possible to check (77) for a *fixed* choice of the coefficients $(a_i)_i$ to ensure that the EPI (8c) holds for these coefficients. Of course, (77) is obviously always satisfied for independent random vectors $(X_i)_i$. In order to relate condition (77) to Takano and Johnson's (74), (75), we rewrite the former in terms of Fisher information as follows.

Corollary 2: For random variables $(X_i)_i$ $(n = 1)$, (77) is equivalent to

$$\mathsf{diag}(J(X_{i,t}))_i \geq \mathbf{J}(X_t) \quad (78)$$

for all $t > 0$, where $X_t = (X_{i,t})_i$. Therefore, if this condition is satisfied then the MII (48) and the EPI (8) hold.

Proof: Let Z be a standard Gaussian random variable independent of X_t, and define $a = (a_i)_i$ and $Y_t = X_t + \sqrt{\varepsilon}\,aZ$, where $aZ = (a_iZ)_i$. The perturbations (76) ensure that the density of X_t is smooth, so that the function $I(\varepsilon) = I\{(Y_{i,t})_i\}$ is differentiable for all $\varepsilon \geq 0$. Now (77) is equivalent to $I(\varepsilon) \geq I(0) + o(\varepsilon)$, that is, $I'(0) \geq 0$. By (51)

$$I\{(Y_{i,t})_i\} = \sum_i h(Y_{i,t}) - h(Y_t)$$

so $I'(0) \geq 0$ can be rewritten as

$$\frac{d}{d\varepsilon}\sum_i h(X_{i,t}+\sqrt{\varepsilon}\,a_iZ) - h(X_t+\sqrt{\varepsilon}\,aZ)\bigg|_{\varepsilon=0} \geq 0.$$

By de Bruijn's identity (36), this is equivalent to

$$\sum_i a_i^2 J(X_{i,t}) \geq \mathsf{tr}(\mathbf{J}(X_t)\mathsf{Cov}(aZ))$$

where $\mathsf{Cov}(aZ) = a\mathsf{Var}(Z)a^t = aa^t$, that is

$$a^t \cdot \mathsf{diag}(J(X_{i;t}))_i \cdot a \geq a^t \cdot \mathbf{J}(X_t) \cdot a \quad (79)$$

for any vector a and $t > 0$. This shows that (77) is equivalent to the matrix inequality (78) as required. □

We now recover Takano and Johnson's conditions (74), (75) from (78).

Lemma 5: In the case of two random variables X_1, X_2, (74) and (75) are equivalent to

$$\lambda^t \cdot \mathsf{diag}(J(X_{i;t}))_i \cdot \lambda \geq \lambda^t \cdot \mathbf{J}(X_t) \cdot \lambda \quad (80)$$

$$\lambda^t \cdot \mathsf{diag}(J(X_{i;t}))_i \cdot \lambda \geq \mu^t \cdot \mathbf{J}(X_t) \cdot \mu \quad (81)$$

respectively, where λ and μ minimize the quadratic forms $a^t \cdot \mathsf{diag}(J(X_{i;t}))_i \cdot a$ and $a^t \cdot \mathbf{J}(X_t) \cdot a$, respectively, over all vectors a of the form $a = (\alpha, 1-\alpha)^t$, $0 \leq \alpha \leq 1$.

Proof: Given a positive definite symmetric matrix $\mathbf{J}$, the general solution a^* to

$$\min_a a^t \cdot \mathbf{J} \cdot a \quad \Big(a_i \geq 0, \sum_i a_i = 1\Big)$$

is easily found by the Lagrangian multiplier method. One finds $a_i^* = \sum_j J_{i,j}^{-1} / \sum_{i,j} J_{i,j}^{-1}$ for all i and $a^{*t} \cdot \mathbf{J} \cdot a^* = (\sum_{i,j} J_{i,j}^{-1})^{-1}$, where $J_{i,j}^{-1}$ are the entries of the inverse matrix $\mathbf{J}^{-1}$. Particularizing this gives $\lambda_1 \propto J^{-1}(X_{1,t})$, $\lambda_2 \propto J^{-1}(X_{2,t})$ and $\mu_1 \propto J_{2,2}(X_t) - J_{1,2}(X_t)$, $\mu_2 \propto J_{1,1}(X_t) - J_{1,2}(X_t)$ up to appropriate proportionality factors, and (80), (81) are rewritten as

$$\begin{aligned} &J^{-1}(X_{1,t}) + J^{-1}(X_{2,t}) \\ &\quad \geq \frac{J_{1,1}(X_t)}{J^2(X_{1,t})} + \frac{J_{2,2}(X_t)}{J^2(X_{2,t})} + 2\frac{J_{1,2}(X_t)}{J(X_{1,t})J(X_{2,t})} \qquad (82) \\ &(J^{-1}(X_{1,t}) + J^{-1}(X_{2,t}))^{-1} \\ &\quad \geq \frac{J_{1,1}(X_t)J_{2,2}(X_t) - J_{1,2}^2(X_t)}{J_{1,1}(X_t) + J_{2,2}(X_t) - 2J_{1,2}(X_t)}. \qquad (83) \end{aligned}$$

Meanwhile, expanding the RHSs in (74), (75) using Stein's identity [42] gives

$$\begin{aligned} &2(\nu_1 + \nu_2) - J^{-1}(X_{1,t}) - J^{-1}(X_{2,t}) \\ &\qquad \geq \nu_1^2 J_{1,1}(X_t) + \nu_2^2 J_{2,2}(X_t) + 2\nu_1\nu_2 J_{1,2}(X_t) \end{aligned}$$

where $(\nu_1, \nu_2) = (J^{-1}(X_{1,t}), J^{-1}(X_{2,t}))$ for Takano's condition and $(\nu_1, \nu_2) = (J_{2,2}(X_t) - J_{1,2}(X_t), J_{1,1}(X_t) - J_{1,2}(X_t))/(J_{1,1}(X_t)J_{2,2}(X_t) - J_{1,2}^2(X_t))$ for Johnson's condition. Replacing yields (82) and (83), respectively. This proves the lemma. □

Corollary 3: In the case of two random variables X_1, X_2, (78) implies both Takano and Johnson's conditions (74), (75).

Proof: Condition (78) implies (79) for any a of the form $a = (\alpha, 1-\alpha)^t$, $0 \leq \alpha \leq 1$. Setting $a = \lambda$ yields Takano's condition (80). Replacing the RHS of the resulting inequality by the minimum over a (achieved by $a = \mu$) gives Johnson's condition (81). □

Thus, our condition (78) is stronger than Takano's or Johnson's. This is not surprising since it yields a stronger form of the EPI (8), valid for any choice of coefficients $(a_i)_i$.

VI. Liu and Viswanath's Covariance-Constrained EPI

As aforementioned in the Introduction, all known applications of the EPI to source and channel coding problems [14]–[26] involve an inequality of the form $N(X+Z) \geq N(X)+N(Z)$, where Z is Gaussian independent of X. In this and the next section, we study generalizations of this inequality. We begin with Liu and Viswanath's generalized EPI for constrained covariance matrices.

A. Background

Recently, Liu and Viswanath [43], [44] have suggested that the EPI's main contribution to multiterminal coding problems is for solving optimization problems of the form

$$\max_{p(x)} h(X) - \mu h(X+Z) \quad (\mu \geq 1) \qquad (84)$$

where Z is Gaussian and the maximization is over all random n-vectors X independent of Z. The solution is easily determined from the EPI in the form (8c) applied to the random vectors $X_1 = \mu^{1/2}X$ and $X_2 = (1-\mu^{-1})^{-1/2}Z$

$$\begin{aligned} h(X+Z) \geq \mu^{-1}&\left(h(X) + \frac{n}{2}\log\mu\right) \\ &+ (1-\mu^{-1})h\left((1-\mu^{-1})^{-1/2}Z\right). \end{aligned}$$

Since equality holds iff X_1 and X_2 are Gaussian with identical covariances, it follows that the optimal solution X to (84) is Gaussian with covariance matrix $\mathsf{Cov}(X) = (\mu-1)^{-1}\mathsf{Cov}(Z)$.

Clearly, the existence of a Gaussian solution to (84) is equivalent to the EPI for two independent random vectors X and Z. Liu and Viswanath [43], [44] have found an implicit generalization of the EPI by showing that (84) still admits a Gaussian solution under the covariance constraint $\mathsf{Cov}(X) \leq \mathbf{C}$, where $\mathbf{C}$ is any positive definite matrix. The gave a "direct proof," motivated by the vector Gaussian broadcast channel problem, using the classical EPI, the saddlepoint property of mutual information (45) and the "enhancement" technique for Gaussian random vectors introduced by Weingarten, Steinberg, and Shamai [15]. They also gave a "perturbation proof" using a generalization of the conventional techniques presented in Section II, namely, an integration over a path of the form $\{\sqrt{1-t}\,X + \sqrt{t}\,Z\}$ of a generalized FII (17c) with matrix coefficients, using de Bruijn's identity and the Cramér-Rao inequality[13]. This and similar results for various optimization problems involving several Gaussian random vectors find applications in vector Gaussian broadcast channels and distributed vector Gaussian source coding [44].

B. An Explicit Covariance-Constrained MII

We first give explicit forms of covariance-constrained MII and EPI, which will be used to solve Liu and Viswanath's optimization problem. Again, the same ideas as in the proof of Theorem 1 are easily generalized to prove the following covariance-constrained MII and EPI, using only basic properties of mutual information.

Theorem 4: Let X_1, X_2 be independent random n-vectors with positive definite covariance matrices, and let Z_1, Z_2 be Gaussian random n-vectors independent of X_1, X_2 and of each other, with covariances proportional to those of X_1 and X_2, respectively: $\mathsf{Cov}(Z_1) = \alpha\mathsf{Cov}(X_1)$, $\mathsf{Cov}(Z_2) = \alpha\mathsf{Cov}(X_2)$, where $\alpha > 0$. Assume that X_2 is Gaussian and X_1, X_2 are subject to the covariance constraint

$$\mathsf{Cov}(X_1) \leq \mathsf{Cov}(X_2). \qquad (85)$$

Then for any real-valued coefficients a_1, a_2 normalized such that $a_1^2 + a_2^2 = 1$

$$\begin{aligned} I(a_1X_1 + a_2X_2 + Z; Z) \leq\ & a_1^2 I(X_1 + Z_1; Z_1) \\ &+ a_2^2 I(X_2 + Z_2; Z_2) \qquad (86) \end{aligned}$$

[13]As explained in Section II-D–7, the Cramér-Rao inequality (46) is equivalent to the saddlepoint property (45) used in their "direct proof."

where we have noted $Z = a_1 Z_1 + a_2 Z_2$. Furthermore, this inequality implies the following generalized EPI:

$$h(a_1 X_1 + a_2 X_2) \geq a_1^2 h(X_1) + a_2^2 h(X_2) + \Delta \tag{87}$$

where

$$\Delta = h(Z) - a_1^2 h(Z_1) - a_2^2 h(Z_2) \geq 0. \tag{88}$$

Note that for the particular case $\mathsf{Cov}(X_1) = \mathsf{Cov}(X_1)$, we have $\mathsf{Cov}(Z_1) = \mathsf{Cov}(Z_2)$, the random vectors Z_1, Z_2 and Z are identically distributed, $\Delta = 0$ and Theorem 4 reduces to Theorem 1 for two random vectors.

Proof of Theorem 4: First define

$$Z_i' = \mathsf{Cov}(Z_i)\mathsf{Cov}(Z)^{-1} Z \quad (i = 1, 2)$$

with covariance matrices $\mathsf{Cov}(Z_i') = \mathsf{Cov}(Z_i)\mathsf{Cov}(Z)^{-1}\mathsf{Cov}(Z_i)$. From (85), one successively has $\mathsf{Cov}(Z_1) \leq \mathsf{Cov}(Z_2)$, $\mathsf{Cov}(Z_1) \leq a_1^2\mathsf{Cov}(Z_1) + a_2^2\mathsf{Cov}(Z_2) = \mathsf{Cov}(Z)$, $\mathsf{Cov}(Z)^{-1} \leq \mathsf{Cov}^{-1}(Z_1)$, and upon left and right multiplication by $\mathsf{Cov}(Z_1)$, $\mathsf{Cov}(Z_1') = \mathsf{Cov}(Z_1)\mathsf{Cov}(Z)^{-1}\mathsf{Cov}(Z_1) \leq \mathsf{Cov}(Z_1)$. Similarly, $\mathsf{Cov}(Z_2') \geq \mathsf{Cov}(Z_2)$. Therefore, we can write

$$\begin{aligned} Z_1 &= Z_1' + \tilde{Z}_1 \\ Z_2' &= Z_2 + \tilde{Z}_2 \end{aligned} \tag{89}$$

where $\tilde{Z}_1$ and $\tilde{Z}_1$ are Gaussian and independent of Z_1' and Z_2, respectively. We can now write the following string of inequalities:

$$\begin{aligned} &I(a_1 X_1 + a_2 X_2 + Z; Z) \\ &\quad = I(a_1(X_1 + a_1 Z_1') + a_2(X_2 + a_2 Z_2'); Z) && \text{(90a)} \\ &\quad \leq I(X_1 + a_1 Z_1', X_2 + a_2 Z_2'; Z) && \text{(90b)} \\ &\quad \leq I(X_1 + a_1 Z_1'; Z_1') + I(X_2 + a_2 Z_2'; Z_2') && \text{(90c)} \\ &\quad = I(X_1 + a_1 Z_1; Z_1) + I(X_2 + a_2 Z_2; Z_2) \\ &\qquad - I(X_1 + a_1(Z_1' + \tilde{Z}_1); \tilde{Z}_1) \\ &\qquad + I(X_2 + a_2(Z_2 + \tilde{Z}_2); \tilde{Z}_2) && \text{(90d)} \end{aligned}$$

where (90a) holds since $a_1^2 Z_1' + a_2^2 Z_2' = \mathsf{Cov}(Z)\mathsf{Cov}(Z)^{-1} Z = Z$, (90b) follows from the data processing theorem applied to the linear transformation (9), (90c) follows from Sato's inequality (Lemma 2), and (90d) follows by applying the identity (58) to the random vectors defined by (89). By Proposition 8, $I(X_1 + a_1(Z_1' + \tilde{Z}_1); \tilde{Z}_1) \geq I(X_1^* + a_1(Z_1' + \tilde{Z}_1); \tilde{Z}_1)$, where X_1^* is Gaussian with covariance matrix $\mathsf{Cov}(X_1^*) = \mathsf{Cov}(X_1)$.

We now use the assumption that $\mathsf{Cov}(Z_1) = \alpha\mathsf{Cov}(X_1)$, $\mathsf{Cov}(Z_2) = \alpha\mathsf{Cov}(X_2)$ and let $\alpha \to 0$ in the well-known expressions for mutual informations of Gaussian random vectors

$$\begin{aligned} &I(X_1^* + a_1(Z_1' + \tilde{Z}_1); \tilde{Z}_1) \\ &\quad = \frac{1}{2}\log\,|\mathbf{I} + \alpha a_1^2 \mathsf{Cov}(\tilde{Z}_1)\mathsf{Cov}(Z_1)^{-1}\,| + o(\alpha) \\ &I(X_2 + a_2(Z_2 + \tilde{Z}_2); \tilde{Z}_2) \\ &\quad = \frac{1}{2}\log\,|\mathbf{I} + \alpha a_2^2 \mathsf{Cov}(\tilde{Z}_2)\mathsf{Cov}(Z_2)^{-1}\,| + o(\alpha) \end{aligned}$$

where $\mathsf{Cov}(\tilde{Z}_1) = \mathsf{Cov}(Z_1) - \mathsf{Cov}(Z_1') = \mathsf{Cov}(Z_1) - \mathsf{Cov}(Z_1)\mathsf{Cov}(Z)^{-1}\mathsf{Cov}(Z_1)$ and $\mathsf{Cov}(\tilde{Z}_2) = \mathsf{Cov}(Z_2') - \mathsf{Cov}(Z_2) = \mathsf{Cov}(Z_2)\mathsf{Cov}(Z)^{-1}\mathsf{Cov}(Z_2) - \mathsf{Cov}(Z_2)$. Since $\mathsf{Cov}(Z) = a_1^2\mathsf{Cov}(Z_1) + a_2^2\mathsf{Cov}(Z_2)$ and $a_1^2 + a_2^2 = 1$, we have $a_1^2\mathsf{Cov}(\tilde{Z}_1)\mathsf{Cov}(Z_1)^{-1} = a_1^2(\mathbf{I} - \mathsf{Cov}(Z_1)\mathsf{Cov}(Z)^{-1}) = a_2^2(\mathsf{Cov}(Z_2)\mathsf{Cov}(Z)^{-1} - \mathbf{I}) = a_2^2\mathsf{Cov}(\tilde{Z}_2)\mathsf{Cov}(Z_2)^{-1}$, and therefore

$$\begin{aligned} &I(X_1 + a_1(Z_1' + \tilde{Z}_1); \tilde{Z}_1) \\ &\quad \geq I(X_1^* + a_1(Z_1' + \tilde{Z}_1); \tilde{Z}_1) \\ &\quad = I(X_2 + a_2(Z_2 + \tilde{Z}_2); \tilde{Z}_2) + o(\alpha). \end{aligned}$$

It follows from (90d) that

$$\begin{aligned} &I(a_1 X_1 + a_2 X_2 + Z; Z) \\ &\quad \leq I(X_1 + a_1 Z_1; Z_1) + I(X_2 + a_2 Z_2; Z_2) \\ &\qquad + o(\alpha). \end{aligned} \tag{92}$$

The rest of the proof is entirely similar to that of Theorem 1. Here is a sketch. Write (92) for $\hat{X}_1 = X_1 + \sqrt{t}\,\hat{Z}_1$ and $\hat{X}_2 = X_2 + \sqrt{t}\,\hat{Z}_2$, where $\hat{Z}_i$ is identically distributed as Z_i and independent of all other random vectors, for $i = 1, 2$. Applying Lemma 3 to the RHS of the resulting inequality, this gives

$$\begin{aligned} &I(a_1 X_1 + a_2 X_2 + \sqrt{t}\hat{Z} + \sqrt{\varepsilon} Z; Z) \\ &\quad \leq a_1^2 I(X_1 + \sqrt{t}\hat{Z}_1 + \sqrt{\varepsilon}\, Z_1; Z_1) \\ &\qquad + a_2^2 I(X_2 + \sqrt{t}\,\hat{Z}_2 + \sqrt{\varepsilon}\, Z_2; Z_2) + o(\varepsilon) \end{aligned}$$

where $\hat{Z}$ is identically distributed as Z. By virtue of (58), this can be written in the form $f(t + \varepsilon) \leq f(t) + o(\varepsilon)$, where

$$\begin{aligned} f(t) = I(a_1 X_1 + a_2 X_2 + \sqrt{t}\, Z; Z) \\ - a_1^2 I(X_1 + \sqrt{t}\, Z_1; Z_1) - a_2^2 I(X_2 + \sqrt{t}\, Z_2; Z_2). \end{aligned}$$

Therefore, $f(t)$ is nonincreasing, and $f(1) \leq f(0)$, which is the required MII (86). This in turn can be rewritten in the form

$$h(a_1 X_1 + a_2 X_2) - a_1^2 h(X_1) - a_2^2 h(X_2) \geq \Delta + \Delta'$$

where Δ is defined by (88) and $\Delta' = I(a_1 X_1 + a_2 X_2; a_1 X_1 + a_2 X_2 + Z) - a_1^2 I(X_1; X_1 + Z_1) - a_2^2 I(X_2; X_2 + Z_2)$ tends to zero as $\alpha \to \infty$ by Lemma 3. This proves (87) and the theorem. □

It is now easy to recover Liu and Viswanath's formulation.

Corollary 4 (Liu and Viswanath [43] and [44]): The maximization problem (84), subject to the covariance constraint $\mathsf{Cov}(X) \leq \mathbf{C}$, admits a Gaussian optimal solution X^*.

Proof: Let X^* be the optimal solution to the maximization problem obtained by restricting the solution space within Gaussian distributions. Thus $\mathsf{Cov}(X^*) > 0$ maximizes

$$\frac{1}{2}\log((2\pi e)^n|\mathsf{Cov}(X)|) - \frac{\mu}{2}\log((2\pi e)^n|\mathsf{Cov}(X) + \mathsf{Cov}(Z)|)$$

over all covariance matrices $\mathsf{Cov}(X) \leq \mathbf{C}$. As stated in [44] and shown in [15], $\mathsf{Cov}(X^*)$ must satisfy the Karush-Kuhn-Tucker condition

$$\frac{1}{2}\mathsf{Cov}(X^*)^{-1} = \frac{\mu}{2}(\mathsf{Cov}(X^*) + \mathsf{Cov}(Z))^{-1} + \mathbf{M}$$

where $\mathbf{M} \geq 0$ is a Lagrange multiplier corresponding to the contsraint $\mathsf{Cov}(X) \leq \mathbf{C}$. It follows that $\mathsf{Cov}(X^*)^{-1} \geq \mu(\mathsf{Cov}(X^*) + \mathsf{Cov}(Z))^{-1}$, that is, $\mu\mathsf{Cov}(X^*) \leq \mathsf{Cov}(X^*) + \mathsf{Cov}(Z)$, or $\mathsf{Cov}(\mu^{1/2}X^*) \leq \mathsf{Cov}((1-\mu^{-1})^{-1/2}Z)$.

Now let X be any random vector independent of Z, such that $\mathsf{Cov}(X) = \mathsf{Cov}(X^*)$. Define $a_1 = \mu^{-1/2}$, $a_2 = (1-\mu^{-1})^{1/2}$, $X_1 = \mu^{1/2}X$, $X_2 = (1-\mu^{-1})^{-1/2}Z$ and $Z_1 = \mu^{1/2}X^*$, and let Z_2 be a Gaussian random vector identically distributed as X_2 and independent of Z_1. Since $a_1^2 + a_2^2 = 1$ and $\mathsf{Cov}(X_1) = \mathsf{Cov}(Z_1) \leq \mathsf{Cov}(X_2) = \mathsf{Cov}(Z_2)$, we may apply Theorem 4. By (87), we obtain

$$h(a_1X_1 + a_2X_2) - a_1^2 h(X_1) \geq h(a_1Z_1 + a_2Z_2) - a_1^2 h(Z_1)$$

that is, replacing and rearranging

$$h(X) - \mu h(X+Z) \leq h(X^*) - \mu h(X^*+Z).$$

Therefore, the Gaussian random vector X^* is an optimal solution to (84) subject to the constraint $\mathsf{Cov}(X) \leq \mathbf{C}$. This completes the proof. □

VII. Costa's EPI: Concavity of Entropy Power

A. Background

Costa [36] has strengthened the EPI for two random vectors X, Z in the case where Z is white Gaussian. While it can be easily shown [4], [36] that Shannon's EPI for X, Z is equivalent to

$$\frac{d}{dt}N(X + \sqrt{t}\,Z) \geq 1$$

Costa's EPI is the convexity inequality which expresses that the entropy power is a concave function of the power of the added Gaussian noise

$$\frac{d^2}{dt^2}N(X + \sqrt{t}\,Z) \leq 0. \tag{93}$$

Alternatively, the concavity of the entropy power is equivalent to saying that the slope $\delta(t) = (N(X + \sqrt{t}\,Z) - N(X))/t$ drawn from the origin is nonincreasing, while the corresponding Shannon's EPI is weaker, being simply equivalent to the inequality $\delta(1) \geq \delta(\infty) = N(Z)$.

The original proof of Costa through an explicit calculation of the second derivative in (93) is quite involved [36]. His calculations are simplified in [107]. Dembo gave an elegant proof using the FII over the path $\{X + \sqrt{t}Z\}$ [4], [37]. Recently, Guo, Shamai, and Verdú provided a clever proof using the MMSE over the path $\{\sqrt{t}X + Z\}$ [12].

Costa's EPI has been used to determine the capacity region of the Gaussian interference channel [20]. It was also used as a continuity argument about entropy that was required for the analysis of the capacity of flat-fading channels in [108].

B. A New Proof of the Concavity of the Entropy Power

In his original presentation [36], Costa proposed the concavity property $N(X + \sqrt{t}\,Z) \geq (1-t)N(X) + tN(X + Z)$ in the segment $(0,1)$ for white Gaussian Z, in which case he showed its equivalence to (93). He also established this inequality in the dual case where X is Gaussian and Z is arbitrary. In the latter case, however, this inequality is not sufficient to prove that $N(X + \sqrt{t}\,Z)$ is a concave function of $t > 0$. In this section, we prove a slight generalization of Costa's EPI, showing concavity in both cases, for an arbitrary (not necessarily white) Gaussian random vector. Again the proposed proof relies only on the basic properties of mutual information.

Theorem 5 (Concavity of Entropy Power): Let X and Z be any two independent random n-vectors. If either X or Z is Gaussian, then $N(X + \sqrt{t}\,Z)$ is a concave function of t.

Proof: To simplify the notation, let $Z_t = \sqrt{t}\,Z$. First, it is sufficient to prove concavity in the case where Z is Gaussian, because, as it is easily checked, the functions $n(t) = N(X+Z_t)$ and $t\cdot n(1/t) = N(X_t+Z)$ are always simultaneously concave. Next define

$$f_X(t) = \frac{N(X+Z_t)}{N(X)} = \exp\frac{2}{n}I(X+Z_t;Z).$$

Our aim is to prove (93), that is, $f_X''(t) \leq 0$. Consider the MII (48) in the form

$$I(X_\lambda + Y_{1-\lambda} + Z_t; Z) \leq \lambda I(X+Z_t;Z) + (1-\lambda)I(Y+Z_t;Z)$$

where Y is independent of X, Z and $0 \leq \lambda \leq 1$. Replacing X_λ, Y_λ by X, Y gives the alternative form

$$I(X+Y+Z_t;Z) \leq \lambda I(X+Z_{\lambda t};Z) + (1-\lambda)I(Y+Z_{(1-\lambda)t};Z).$$

Choose Z' and Z'' such that Z, Z' and Z'' are i.i.d. and independent of X, and replace X by $X + Z_u'$ and Y by Z_v''. This gives

$$\begin{aligned} &I(X + Z_u' + Z_v'' + Z_t; Z) \\ &\quad\leq \lambda I(X + Z_u' + Z_{\lambda t}; Z) \\ &\qquad + (1-\lambda) I(Z_v'' + Z_{(1-\lambda)t}; Z). \end{aligned} \tag{94}$$

We now turn this into a "mutual information power inequality" similarly as the EPI (8a) is derived from (8c) in the proof of Proposition 2. Define $M_t(X)$ as the power of a Gaussian random vector $\tilde{X}$ having covariances proportional to those of Z and identical mutual information $I(X+Z_t;Z)$. By Shannon's capacity formula, $I(\tilde{X}+Z_t;Z) = \frac{n}{2}\log(1 + t\sigma_Z^2/\sigma_{\tilde{X}}^2)$, and therefore

$$M_t(X) = \frac{t\sigma_Z^2}{\exp\frac{2}{n}I(X+Z_t;Z) - 1} = \frac{t\sigma_Z^2}{f_X(t) - 1}.$$

Choose $\lambda \in [0,1]$ such that $I(X + Z_u' + Z_{\lambda t}; Z) = I(Z_v'' + Z_{(1-\lambda)t}; Z)$ in (94). This is always possible, because the difference has opposite signs for $\lambda = 0$ and $\lambda = 1$. By applying the function $(\exp(\frac{2}{n}\cdot) - 1)^{-1}$ to both sides of (94), we find

$$M_t(X + Z_u' + Z_v'') \geq M_{\lambda t}(X + Z_u') + v\sigma_Z^2.$$

We now let $t \to 0$ (so that $\lambda t \to 0$). Since $f_{X+Z_u'}(t) = f_X(t+u)/f_X(u)$, and similarly, $f_{X+Z_u'+Z_v''}(t) = f_X(t+u+v)/f_X(u+v)$, we obtain

$$\frac{f_X(u+v)}{f_X'(u+v)} \geq \frac{f_X(u)}{f_X'(u)} + v.$$

Dividing by v and letting $v \to 0$ gives

$$\frac{d}{du}\frac{f_X(u)}{f'_X(u)} \geq 1,$$

that is, carrying out the derivation, $f_X(u)f''_X(u)/f'_X(u)^2 \leq 0$ or $f''_X(u) \leq 0$ as required. □

It would be interesting to know whether this proof can be adapted to the recent generalization of Costa's EPI [109]–[111] in which t is replaced by an arbitrary positive semidefinite matrix.

VIII. Open Questions

A. EPI, FII, and MII for Subsets of Independent Variables

Recently, Artstein, Ball, Barthe, and Naor [45] proved a new entropy power inequality involving entropy powers of sums of all independent variables excluding one, which solved a long-standing conjecture about the monotonicity of entropy. This was generalized to arbitrary collections of subsets of independent variables (or vectors) by Madiman and Barron [46], [47]. The generalization of the classical formulation of the EPI takes the form

$$N\big(\sum_i X_i\big) \geq \frac{1}{k}\sum_S N\big(\sum_{i\in S} X_i\big) \tag{95}$$

where the sum in the RHS is over arbitrary subsets S of indexes, and k is the maximum number of subsets in which one index appears. Note that we may always assume that subsets S are "balanced" [47], i.e., each index i appears in the RHS of (95) exactly k times. This is because it is always possible to add singletons to a given collection of subsets until the balancing condition is met; since the EPI (95) would hold for the augmented collection, it a fortiori holds for the initial collection as well.

For balanced subsets, the inequalities generalizing (8c), (17c), (33), and (48) are the following.

Proposition 10: Let $(X_i)_i$ be finitely many random n-vectors, let Z be any Gaussian random n-vector independent of $(X_i)_i$, and let $(a_i)_i$ be any real-valued coefficients normalized such that $\sum_i a_i^2 = 1$. Then, for any collection $\{S\}$ of balanced subsets of indexes

$$J\big(\sum_i a_i X_i\big) \leq \sum_S a_S^2 J(X_S), \tag{96a}$$

$$\mathsf{Var}\big(\sum_i a_i X_i \,|\, \sum_i a_i X_i + Z\big) \geq \sum_S a_S^2 \mathsf{Var}(X_S \,|\, X_S + Z) \tag{96b}$$

$$h\big(\sum_i a_i X_i\big) \geq \sum_S a_S^2\, h(X_S) \tag{96c}$$

$$I\big(\sum_i a_i X_i + Z; Z\big) \leq \sum_S a_S^2\, I(X_S + Z; Z) \tag{96d}$$

where $a_S^2 = \frac{1}{k}\sum_{i\in S} a_i^2$ (so that $\sum_S a_S^2 = 1$) and X_S is given by the covariance preserving transformation

$$X_S = \frac{\sum_{i\in S} a_i X_i}{\sqrt{\sum_{i\in S} a_i^2}}.$$

Available proofs of (96a)–(96c) are generalizations of the conventional techniques presented in Section II, where an additional tool ("variance drop lemma") is needed to prove either (96a) or (96b); see [45, Lemma 5], [48, Lemma 3], or [47, Lemma 2]. Artsein, Ball, Barthe, and Naor's proof of the EPI (96c), which is generalized and simplified by Madiman and Barron, is through an integration of the FII (96a) over the path $\{\sqrt{t}\,X + \sqrt{1-t}\,Z\}$ (in [45, eq. (47b)]) or $\{X + \sqrt{t}\,Z\}$ (in [47, eq. (47a)]). Tulino and Verdú provided the corresponding proof via MMSE [48], through an integration of the MMSE inequality (96b) over the path $\{\sqrt{t}\,X + Z\}$ [see (47c)]. Again the approaches corresponding to (96a) and (96b) are equivalent by virtue of the complementary relation (29), as explained in Section II-C–3.

That the MII (96d) holds is easily shown through (96a) or (96b) and de Bruijn's identity (41) or (42). However, the author was not able to extend the ideas in the proof of Theorem 1 to provide a direct proof of the MII (96d), which letting $\sigma_Z^2 \to \infty$ would yield an easy proof of the generalized EPI (96c). Such an extension perhaps involves a generalization of the data processing inequality or Sato's inequality in (59), which using the relation $\sum_i a_i X_i = \frac{1}{\sqrt{k}}\sum_S a_S X_S$ would yield the inequality $I(\sum_i a_i X_i + Z; Z) \leq \sum_S I(X_S + a_s Z; Z) + o(\sigma_Z^2)$.

B. EPI, FII, and MII for Gas Mixtures

There is a striking resemblance between the original inequalities (8c), (17c), (33), and (48) for linear mixtures of independent random vectors, and known inequalities concerning entropy and Fisher information for linear "gas mixtures" of probability distributions.

Proposition 11: Let the random variable I have distribution $p(i) = a_i^2$ where $\sum_i a_i^2 = 1$, let $(X_i)_i$ be finitely many random n-vectors independent of I, and let Z be white Gaussian, independent of $(X_i)_i$ and I. Then

$$J(X_I) \leq \sum_i a_i^2 J(X_i) \tag{97a}$$

$$\mathsf{Var}(X_I \,|\, X_I + Z) \geq \sum_i a_i^2 \mathsf{Var}(X_i \,|\, X_i + Z) \tag{97b}$$

$$h(X_I) \geq \sum_i a_i^2 h(X_i) \tag{97c}$$

$$I(X_I + Z; Z) \leq \sum_i a_i^2 I(X_i + Z; Z). \tag{97d}$$

Noting that X_I has distribution $p_{X_I}(x) = \sum_i p(i)p(x\,|\,i) = \sum_i a_i^2 p_{X_i}(x)$, the "FII" (97a) can be proved directly as follows. Let $S(x) = \nabla p_{X_I}(x)/p_{X_I}(x)$ and $S_i(x) = \nabla p_{X_i}(x)/p_{X_i}(x)$ be the score functions of X_I and the $(X_i)_i$, respectively, and define $\lambda_i(x) = a_i^2 p_{X_i}(x)/p_{X_I}(x)$ for all i. Then $\sum_i \lambda_i(x) = 1$, $S(x) = \sum_i \lambda_i(x) S_i(x)$ and since the squared norm is convex, $\|S(x)\|^2 \leq \sum_i \lambda_i(x)\|S_i(x)\|^2 = p_{X_I}^{-1}(x)\sum_i a_i^2 p_{X_i}(x)\|S_i(x)\|^2$. Averaging over $p_{X_I}(x)$ gives (97a).

Once (97a) is established, the conventional techniques presented in Section II can be easily adapted to deduce the other inequalities (97b)–(97d): Substituting $(X_i + Z_i)_i$ for $(X_i)_i$ in (97a), where the $(Z_i)_i$ are independent copies of Z, and noting that $X_I + Z_I$ has the same probability distribution as $X_I + Z$,

we obtain the inequality $J(X_I + Z) \leq \sum_i a_i^2 J(X_i + Z)$; applying the complementary relation (29) gives (97b); integrating using de Bruijn's identity (41) or (42) gives (97d), from which (97c) follows as in the proof of Theorem 1.

In the present case, however, (97c) and (97d) are already well known. In fact, since $p_{X_I}(x) = \sum_i a_i^2 p_{X_i}(x)$ is a convex combination of distributions, the "EPI" (97c) is nothing but the classical concavity property of entropy, seen as a functional of the probability distribution [9], [87], [88]. This is easily established by noting that since conditioning decreases entropy, $h(X_I) \geq h(X_I \,|\, I) = \sum_i p(i) h(X_i)$. Also the "MII" (97d) is just the classical convexity of mutual information $I(Y, Z)$, seen as a functional of the distribution $p(y \,|\, z)$ for fixed $p(z)$ [9, Thm. 2.7.4], [87, Thm. 4.4.3], [88, Thm. 1.7].

Accordingly, we may reverse the order of implication and derive the corresponding convexity property of Fisher information (97a) anew from the "MII" (97d). Indeed, (97d) can be rewritten in the form

$$h(X_I + \sqrt{t}\, Z) - h(X_I) \leq \sum_i a_i^2 (h(X_i + \sqrt{t}\, Z) - h(X_i))$$

for any $t > 0$. Dividing both sides by t and letting $t \to 0$ gives (97a) by virtue of de Bruijn's identity. This derivation is much shorter than earlier proofs of (97a), [4, Lemma 6], [112]. The convexity property of Fisher information finds application in channel estimation [56] and thermodynamics [113].

ACKNOWLEDGMENT

The author wishes to thank the Associate Editor for his patience, and the anonymous reviewers for their helpful comments and for pointing out several references and the counterexample of Section I-B. Preliminary advice from Prof. S. Shamai, Prof. J. Chen, and Dr. Y. Wu are also gratefully acknowledged.

REFERENCES

[1] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 623–656, Oct. 1948.
[2] A. J. Stam, "Some inequalities satisfied by the quantities of information of Fisher and Shannon," *Inf. Contr.*, vol. 2, pp. 101–112, Jun. 1959.
[3] N. M. Blachman, "The convolution inequality for entropy powers," *IEEE Trans. Inf. Theory*, vol. 11, no. 2, pp. 267–271, Apr. 1965.
[4] A. Dembo, T. M. Cover, and J. A. Thomas, "Information theoretic inequalities," *IEEE Trans. Inf. Theory*, vol. 37, no. 6, pp. 1501–1518, Nov. 1991.
[5] E. A. Carlen and A. Soffer, "Entropy production by block variable summation and central limit theorems," *Commun. Math. Phys.*, vol. 140, no. 2, pp. 339–371, 1991.
[6] E. H. Lieb, "Proof of an entropy conjecture of Wehrl," *Commun. Math. Phys.*, vol. 62, pp. 35–41, 1978.
[7] M. H. M. Costa and T. M. Cover, "On the similarity of the entropy power inequality and the Brunn-Minkowski inequality," *IEEE Trans. Inf. Theory*, vol. 30, no. 6, pp. 837–839, Nov. 1984.
[8] O. G. Guleryuz, E. Lutwak, D. Yang, and G. Zhang, "Information-theoretic inequalities for contoured probability distributions," *IEEE Trans. Inf. Theory*, vol. 48, no. 8, pp. 2377–2383, Aug. 2002.
[9] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. New York: Wiley, 2006.
[10] D. Guo, S. Shamai (Shitz), and S. Verdú, "Mutual information and minimum mean-square error in Gaussian channels," *IEEE Trans. Inf. Theory*, vol. 51, no. 4, pp. 1261–1282, Apr. 2005.
[11] S. Verdú and D. Guo, "A simple proof of the entropy-power inequality," *IEEE Trans. Inf. Theory*, vol. 52, no. 5, pp. 2165–2166, May 2006.
[12] D. Guo, S. Shamai (Shitz), and S. Verdú, "Proof of entropy power inequalities via MMSE," in *Proc. IEEE Int. Symp. Inf. Theory*, Seattle, WA, July 2006, pp. 1011–1015.
[13] J. Binia, "On divergence-power inequalities," *IEEE Trans. Inf. Theory*, vol. 53, no. 3, pp. 1179–1182, Mar. 2007.
[14] P. Bergmans, "A simple converse for broadcast channels with additive white Gaussian noise," *IEEE Trans. Inf. Theory*, vol. 20, no. 2, pp. 279–280, Mar. 1974.
[15] H. Weingarten, Y. Steinberg, and S. Shamai (Shitz), "The capacity region of the Gaussian multiple-input multiple-output broadcast channel," *IEEE Trans. Inf. Theory*, vol. 52, no. 9, pp. 3936–3964, Sep. 2006.
[16] M. Mohseni and J. M. Cioffi, "A proof of the converse for the capacity of Gaussian MIMO broadcast channels," in *Proc. IEEE Int. Symp. Inf. Theory*, Seattle, WA, Jul. 2006, pp. 881–885.
[17] S. Leung-Yan-Cheong and M. Hellman, "The Gaussian wire-tap channel," *IEEE Trans. Inf. Theory* , vol. 24, no. 4, pp. 451–456, Jul. 1978.
[18] E. Tekin and A. Yener, "The Gaussian multiple access wire-tap channel with collective secrecy constraints," in *Proc. IEEE Int. Symp. Inf. Theory*, Seattle, WA, Jul. 2006, pp. 1164–1168.
[19] E. Tekin and A. Yener, "The Gaussian multiple access wire-tap channel," *IEEE Trans. Inf. Theory*, vol. 54, no. 12, pp. 5747–5755, Dec. 2008.
[20] M. H. M. Costa, "On the Gaussian interference channel," *IEEE Trans. Inf. Theory*, vol. 31, no. 5, pp. 607–615, Sep. 1985.
[21] L. Ozarow, "On a source-coding problem with two channels and three receivers," *Bell Syst. Tech. J.*, vol. 59, no. 10, pp. 1909–1921, Dec. 1980.
[22] R. Zamir, "Gaussian codes and Shannon bounds for multiple descriptions," *IEEE Trans. Inf. Theory*, vol. 45, no. 7, pp. 2629–2636, Nov. 1999.
[23] Y. Oohama, "Gaussian multiterminal source coding," *IEEE Trans. Inf. Theory* , vol. 43, no. 6, pp. 1912–1923, Nov. 1997.
[24] Y. Oohama, "The rate-distortion function for the quadratic Gaussian CEO problem," *IEEE Trans. Inf. Theory*, vol. 44, no. 3, pp. 1057–1070, May 1998.
[25] Y. Oohama, "Rate-distortion theory for Gaussian multiterminal source coding systems with several side informations at the decoder," *IEEE Trans. Inf. Theory*, vol. 51, no. 7, pp. 2577–2593, Jul. 2005.
[26] Y. Oohama, "Gaussian multiterminal source coding with several side informations at the decoder," in *Proc. IEEE Int. Symp. Inf. Theory*, Seattle, WA, Jul. 2006, pp. 1409–1413.
[27] D. Donoho, "On minimum entropy deconvolution," in *Applied Time Series Analysis II*. New York: Academic, 1981, pp. 565–608.
[28] J. F. Bercher and C. Vignat, "Estimating the entropy of a signal with applications," *IEEE Trans. Signal Process.*, vol. 48, no. 6, pp. 1687–1694, Jun. 2000.
[29] F. Vrins and M. Verleysen, "On the entropy minimization of a linear mixture of variables for source separation," *Signal Process.*, vol. 85, no. 5, pp. 1029–1044, May 2005.
[30] A. R. Barron, "Entropy and the central limit theorem," *Ann. Probabil.*, vol. 14, no. 1, pp. 336–342, 1986.
[31] L. D. Brown, "A proof of the central limit theorem motivated by the Cramér-Rao inequality," in *Statistics and Probability: Essays in Honor of C. R. Rao*, G. Kallianpur, P. R. Krishnaiah, and J. K. Gosh, Eds. Amsterdam: North-Holland, 1982, pp. 141–148.
[32] O. Johnson, "Entropy inequalities and the central limit theorem;," *Stochastic Process. Their Appl.*, vol. 88, no. 2, pp. 291–304, 2000.
[33] O. Johnson and Y. Suhov, "Entropy and random vectors," *J. Statist. Phys.*, vol. 104, no. 112, pp. 145–165, 2001.
[34] S. Artstein, K. M. Ball, F. Barthe, and A. Naor, "On the rate of convergence in the entropic central limit theorem," *Probabil. Theory Rel. Fields*, vol. 129, no. 3, pp. 381–390, Jul. 2004.
[35] O. Johnson and A. Barron, "Fisher information inequalities and the central limit theorem," *Probabil. Theory Rel. Fields*, vol. 129, no. 3, pp. 391–409, July 2004.
[36] M. H. M. Costa, "A new entropy power inequality," *IEEE Trans. Inf. Theory*, vol. 31, no. 6, pp. 751–760, Nov. 1985.
[37] A. Dembo, "Simple proof of the concavity of the entropy power with respect to added Gaussian noise," *IEEE Trans. Inf. Theory*, vol. 35, no. 4, pp. 887–888, Jul. 1989.
[38] R. Zamir and M. Feder, "A generalization of the entropy power inequality with applications," *IEEE Trans. Inf. Theory*, vol. 39, no. 5, pp. 1723–1728, Sep. 1993.

[39] R. Zamir and M. Feder, "A generalization of the entropy power inequality with applications to linear transformation of a white-noise," in *Proc. IEEE Int. Symp. Inf. Theory*, Jan. 1993, p. 74.

[40] R. Zamir and M. Feder, "A generalization of information theoretic inequalities to linear transformations of independent vector," in *Proc. 6th Joint Swedish-Russian Int. Workshop on Inf. Theory*, Mölle, Sweden, Aug. 1993, pp. 254–258.

[41] S. Takano, S. Watanabe, M. Fukushima, Y. Prohorov, and A. Shiryaev, Eds., "The inequalities of Fisher information and entropy power for dependent variables," in *Proc. 7th Japan-Russia Symp. Probabil. Theory and Math. Statist.*, Tokyo, Jul. 26–30, 1995, Singapore: World Scientific, 1996,pp. 460-470.

[42] O. Johnson, "A conditional entropy power inequality for dependent variables," *IEEE Trans. Inf. Theory*, vol. 50, no. 8, pp. 1581–1583, Aug. 2004.

[43] T. Liu and P. Viswanath, "An extremal inequality motivated by multiterminal information theoretic problems," in *Proc. IEEE Int. Symp. Inf. Theory*, Seattle, WA, Jul. 2006, pp. 1016–1020.

[44] T. Liu and P. Viswanath, "An extremal inequality motivated by multiterminal information theoretic problems," *IEEE Trans. Inf. Theory*, vol. 53, no. 5, pp. 1839–1851, May 2007.

[45] S. Artstein, K. M. Ball, F. Barthe, and A. Naor, "Solution of Shannon's problem on the monotonicity of entropy," *J. Amer. Math. Soc.*, vol. 17, no. 4, pp. 975–982, May 2004.

[46] M. Madiman and A. Barron, "The monotonicity of information in the central limit theorem and entropy power inequalities," in *Proc. IEEE Int. Symp. Inf. Theory*, Seattle, WA, Jul. 2006, pp. 1021–1025.

[47] M. Madiman and A. Barron, "Generalized entropy power inequalities and monotonicity properties of information," *IEEE Trans. Inf. Theory*, vol. 53, no. 7, pp. 2317–2329, Jul. 2007.

[48] A. M. Tulino and S. Verdú, "Monotonic decrease of the non-Gaussianness of the sum of independent random variables: A simple proof," *IEEE Trans. Inf. Theory*, vol. 52, no. 9, pp. 4295–4297, Sep. 2006.

[49] I. Vajda, *Theory of Statistical Inference and Information*. Boston, MA: Kluwer Academic, 1989.

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

[51] O. Johnson, *Information Theory and The Central Limit Theorem*. London, U.K.: Imperial College Press, 2004.

[52] V. Papathanasiou, "Some characteristic properties of the Fisher information matrix via Cacoullos-type inequalities," *Journal of Multivariate Analysis*, vol. 44, no. 2, pp. 256–265, Feb. 1993.

[53] R. Zamir, "A proof of the Fisher information inequality via a data processing argument," *IEEE Trans. Inf. Theory*, vol. 44, no. 3, pp. 1246–1250, May 1998.

[54] Y. Itoh, "An application of the convolution inequality for the Fisher information," *Ann. Inst. Statist. Math.*, vol. 41, no. 1, pp. 9–12, 1989.

[55] E. L. Lehmann and G. Casella, *Theory of Point Estimation*, 2nd ed. New York: Springer, 1998.

[56] C. Budianu and L. Tong, "Channel estimation under asynchronous packet interference," *IEEE Trans. Signal Process.*, vol. 53, no. 1, pp. 333–345, Jan. 2005.

[57] D. Guo, S. Shamai (Shitz), and S. Verdú, "Mutual information and MMSE in Gaussian channels," in *Proceedings of the IEEE Int. Symp. Inf. Theory*, Chicago, IL, Jun. 2004, p. 347.

[58] D. P. Palomar and S. Verdú, "Gradient of mutual information in linear vector Gaussian channels," *IEEE Trans. Inf. Theory*, vol. 52, no. 1, pp. 141–154, Jan. 2006.

[59] M. S. Pinsker, V. V. Prelov, and E. C. van der Meulen, "Information rates in certain stationary non-Gaussian channels in weak-signal transmission," *Problems of Inf. Transmiss.*, vol. 34, no. 1, pp. 1–13, 1998.

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

[61] J. M. Borden, D. M. Mason, and R. J. McEliece, "Some information theoretic saddlepoints," *SIAM J. Control Optimiz.*, vol. 23, no. 1, pp. 129–143, Jan. 1985.

[62] S. N. Diggavi and T. M. Cover, "The worst additive noise under a covariance constraint," *IEEE Trans. Inf. Theory*, vol. 47, no. 7, pp. 3072–3081, Nov. 2001.

[63] A. Lapidoth and S. Shamai (Shitz), "Fading channels: How perfect need "perfect side information" be?," *IEEE Trans. Inf. Theory*, vol. 48, no. 5, pp. 1118–1134, May 2002.

[64] D. Guo, S. Shamai (Shitz), and S. Verdu, "Additive non-Gaussian noise channels: Mutual information and conditional mean estimation," in *Proc. IEEE Int. Symp. Inf. Theory*, Adelaide, Australia, Sep. 2005, pp. 719–723.

[65] Y. V. Linnik, "An information-theoretic proof of the central limit theorem with the Lindeberg condition," *Theory of Probabil. Its Appl.*, vol. 4, no. 3, pp. 288–299, 1959.

[66] Y. Itoh, "The information theoretic proof of Kac's theorem," *Proc. Japan Acad.*, vol. 46, no. 3, pp. 283–286, 1970.

[67] V. V. Prelov, "Asymptotic behavior of the capacity of a continuous channel with a large amount of noise," *Problems of Inf. Transmiss.*, vol. 6, no. 2, pp. 122–135, 1970.

[68] V. V. Prelov, "Asymptotic behavior of the capacity of a continuous channel with large nonadditive noise," *Problems of Inf. Transmiss.*, vol. 8, no. 4, pp. 285–289, 1972.

[69] I. A. Ibragimov and R. Z. Khas'minskii, "Weak signal transmission in a memoryless channel," *Problems of Inf. Transmiss.*, vol. 8, no. 4, pp. 290–299, 1972.

[70] V. V. Prelov, "Communication channel capacity with almost Gaussian noise," *Theory of Probabil. Its Appl.*, vol. 33, no. 3, pp. 405–422, Jul. 1988.

[71] V. V. Prelov, "Asymptotic expansions for the mutual information and for the capacity of continuous memoryless channels with weak input signal," *Problems of Contr. Inf. Theory*, vol. 18, no. 2, pp. 91–106, 1989.

[72] V. V. Prelov and E. C. van der Meulen, "An asymptotic expression for the information and capacity of a multidimensional channel with weak input signals," *IEEE Trans. Inf. Theory*, vol. 39, no. 5, pp. 1728–1735, Sept. 1993.

[73] M. S. Pinsker, V. V. Prelov, and S. Verdú, "Sensitivity of channel capacity," *IEEE Trans. Inf. Theory*, vol. 41, no. 6, pp. 1877–1888, Nov. 1995.

[74] M. S. Pinsker, V. V. Prelov, and E. C. van der Meulen, "Weak signal transmission over certain stationary non-Gaussian channels," in *Proc. IEEE Int. Symp. Inf. Theory*, Ulm, Germany, Jun. 1997, p. 371.

[75] M. S. Pinsker, V. V. Prelov, and E. C. van der Meulen, "Information transmission over channels with additive-multiplicative noise," in *Proc. IEEE Int. Symp. Inf. Theory*, Cambridge, MA, Aug. 1998, p. 239.

[76] V. V. Prelov and E. C. van der Meulen, "Higher order asymptotics of mutual information for nonlinear channels with non-Gaussian noise," *Problems of Inf. Transmiss.*, vol. 39, no. 4, pp. 324–340, 2003.

[77] V. V. Prelov and E. C. van der Meulen, "Higher order asymptotics of mutual information for nonlinear channels with non-Gaussian noise," in *Proc. IEEE Int. Symp. Inf. Theory*, Yokohama, Japan, Jun. 2003, p. 83.

[78] V. V. Prelov and S. Verdú, "Second-order asymptotics of mutual information," *IEEE Trans. Inf. Theory*, vol. 50, no. 8, pp. 1567–1580, Aug. 2004.

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

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

[81] D.-T. Pham, "Entropy of a variable slightly contaminated with another," *IEEE Signal Process. Lett.*, vol. 12, no. 7, pp. 536–539, Jul. 2005.

[82] D.-T. Pham and F. Vrins, "Local minima of information-theoretic criteria in blind source separation," *IEEE Signal Process. Lett.*, vol. 12, no. 11, pp. 788–791, Nov. 2005.

[83] F. Vrins, D.-T. Pham, and M. Verleysen, "Mixing and non-mixing local minima of the entropy contrast for blind source separation," *IEEE Trans. Inf. Theory*, vol. 53, no. 3, pp. 1030–1042, Mar. 2007.

[84] S. Ihara, "On the capacity of channels with additive non-Gaussian noise," *Inf. Contr.*, vol. 37, no. 1, pp. 34–39, Apr. 1978.

[85] M. S. Pinsker, "Calculation of the rate of information production by means of stationary random processes and the capacity of stationary channel," *Doklady Akademii nauk SSSR*, vol. 111, pp. 753–756, 1956, (in Russian).

[86] H. Sato, "An outer bound to the capacity region of broadcast channels," *IEEE Trans. Inf. Theory*, vol. 24, no. 3, pp. 374–377, May 1978.

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

[88] R. J. McEliece, *The Theory of Information and Coding*, ser. Encyclopedia of Mathematics and its Applications, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2002.

[89] E. C. Posner, "Random coding strategies for minimum entropy," *IEEE Trans. Inf. Theory*, vol. 21, no. 4, pp. 388–391, July 1975.

[90] G. Caire and S. Shamai, "On the achievable throughput of a multiantenna Gaussian broadcast channel," *IEEE Trans. Inf. Theory*, vol. 49, no. 7, pp. 1691–1706, Jul. 2003.

[91] H. Wang and P. Viswanath, "Vector Gaussian multiple description with individual and central receivers," *IEEE Trans. Inf. Theory*, vol. 53, no. 6, pp. 2133–2153, June 2007.
[92] A. D. Wyner and J. Ziv, "A theorem on the entropy of certain binary sequences and applications: Part I," *IEEE Trans. Inf. Theory*, vol. 19, no. 6, pp. 769–772, Nov. 1973.
[93] A. D. Wyner, "A theorem on the entropy of certain binary sequences and applications: Part II," *IEEE Trans. Inf. Theory*, vol. 19, no. 6, pp. 772–777, Nov. 1973.
[94] H. S. Witsenhausen, "Entropy inequalities for discrete channels," *IEEE Trans. Inf. Theory*, vol. 20, no. 5, pp. 610–616, Sep. 1974.
[95] S. Shamai (Shitz) and A. D. Wyner, "A binary analog to the entropy-power inequality," *IEEE Trans. Inf. Theory*, vol. 36, no. 6, pp. 1428–1430, Nov. 1990.
[96] A. Kagan, "A discrete version of the Stam inequality and a characterization of the Poisson distribution," *J. Statist. Plann. Inference*, vol. 92, pp. 7–12, 2001.
[97] I. Kontoyiannis, P. Harremoës, and O. Johnson, "Entropy and the law of small numbers," *IEEE Trans. Inf. Theory*, vol. 51, no. 2, pp. 466–472, Feb. 2005.
[98] P. Harremoës and C. Vignat, "An entropy power inequality for the binomial family," *J. Inequal. Pure Appl. Math.*, vol. 4, no. 5, 2003.
[99] Y. Yu and O. Johnson, "Concavity of entropy under thinning," in *Proc. IEEE Int. Symp. Inf. Theory*, Seoul, Korea, Jun. 28–Jul. 3 2009, pp. 144–148.
[100] R. Zamir and M. Feder, "Rate-distortion performance in coding bandlimited sources by sampling and dithered quantization," *IEEE Trans. Inf. Theory*, vol. 41, no. 1, pp. 141–154, Jan. 1995.
[101] R. Zamir and M. Feder, "A matrix form of the Brunn-Minkowski inequality," in *Proc. IEEE Int. Symp. Inf. Theory*, Sep. 1995, p. 71.
[102] R. Zamir and M. Feder, "On the volume of the Minkowski sum of line sets and the entropy-power inequality," *IEEE Trans. Inf. Theory*, vol. 44, no. 7, pp. 3039–3043, Nov. 1998.
[103] C. Vignat and J. F. Bercher, "Matrix Fisher inequalities for non-invertible linear systems," in *Proc. IEEE Int. Symp. Inf. Theory*, Lausanne, Switzerland, Jul. 2002, p. 237.
[104] C. Vignat and J. F. Bercher, "On Fisher information inequalities and score functions in non-invertible linear systems," *J. Inequal. Pure and Appl. Math.*, vol. 4, no. 4, pp. 1–9, 2003, Article 71.
[105] R. Zamir, A Necessary and Sufficient Condition for Equality in the Matrix Fisher-Information-Inequality Dept. Elect. Eng. Syst., Tel Aviv Univ., Israel, Tech. Rep., Nov. 1997.
[106] C. F. Caiafa, E. E. Kuruoğlu, and A. N. Proto, "A minimax entropy method for blind separation of dependent components in astrophysical images," in *Proc. 26th Int. Workshop on Bayesian Inference and Max. Entropy Methods in Sci. Eng. CNRS MaxEnt2006*, Paris, France, Jul. 2006.
[107] C. Villani, "A short proof of the "concavity of entropy power"," *IEEE Trans. Inf. Theory*, vol. 46, no. 4, pp. 1695–1696, July 2000.
[108] A. Lapidoth and S. M. Moser, "Capacity bounds via duality with applications to multiple-antenna systems on flat-fading channels," *IEEE Trans. Inf. Theory*, vol. 49, no. 10, pp. 2426–2467, Oct. 2003.
[109] M. Payaró and D. P. Palomar, "A multivariate generalization of Costa's entropy power inequality," in *Proc. IEEE Int. Symp. Inf. Theory*, Toronto, Canada, Jul. 6–11, 2008, pp. 1088–1092.
[110] M. Payaró and D. P. Palomar, "Hessian and concavity of mutual information, differential entropy, and entropy power in linear vector gaussian channels," *IEEE Trans. Inf. Theory*, vol. 55, no. 8, pp. 3613–3628, Aug. 2009.
[111] R. Liu, T. Liu, H. V. Poor, and S. Shamai (Shitz), "A vector generalization of Costa's entropy-power inequality with applications," *IEEE Trans. Inf. Theory*, vol. 56, no. 4, pp. 1865–1879, Apr. 2010.
[112] C. Budianu and L. Tong, Channel Estimation Under Asynchronous Packet Interference Cornell Univ., Ithaca, NY, Tech. Rep. ACSP TR-03-03-01, Mar. 2003.
[113] B. R. Frieden, A. Plastino, A. R. Plastino, and B. H. Soffer, "Fisher-based thermodynamics: Its Legendre transform and concavity properties," *Phys. Rev. E*, vol. 60, no. 1, pp. 48–53, Jul. 1999.

Olivier Rioul (M'07) was born on July 4, 1964. He received the Dipl. Ing. degree from the Ecole Polytechnique, Palaiseau, France, in 1987, and the Dipl. Ing. and the Ph.D. degrees in electrical engineering from the Ecole Nationale Supérieure des Télécommunications (ENST), Paris, France, in 1989 and 1993, respectively.

From 1989 to 1994, he was with the Centre National d'Etudes des Télécommunications (CNET), Issy-les-Moulineaux, France, where he worked on wavelet theory and image compression. In 1994, he joined the ENST, now Télécom ParisTech, where he is currently Associate Professor. His research interests include joint source-channel coding and information theory.