

## A Generalization of the Entropy Power Inequality with Applications

Ram Zamir and Meir Feder, *Senior Member, IEEE*

**Abstract**—We prove the following generalization of the Entropy Power Inequality:

$$h(A\mathbf{x}) \geq h(A\tilde{\mathbf{x}})$$

where  $h(\cdot)$  denotes (joint-) differential-entropy,  $\mathbf{x} = x_1 \dots x_n$  is a random vector with independent components,  $\tilde{\mathbf{x}} = \tilde{x}_1 \dots \tilde{x}_n$  is a Gaussian vector with independent components such that  $h(\tilde{x}_i) = h(x_i)$ ,  $i = 1 \dots n$ , and  $A$  is any matrix. This generalization of the entropy-power inequality is applied to show that a non-Gaussian vector with independent components becomes “closer” to Gaussianity after a linear transformation, where the distance to Gaussianity is measured by the information divergence. Another application is a lower bound, greater than zero, for the mutual-information between nonoverlapping spectral components of a non-Gaussian white process. Finally, we describe a dual generalization of the Fisher Information Inequality.

**Index Terms**—Entropy power inequality, non-Gaussianity, divergence, Fisher information inequality.

### I. THE GENERALIZATION OF THE ENTROPY POWER INEQUALITY

Consider the (joint-) differential-entropy  $h(A\mathbf{x})$  of a linear transformation  $\mathbf{y} = A\mathbf{x}$ , where  $\mathbf{x} = x_1 \dots x_n$  is a vector and

$$h(\mathbf{y}) \triangleq E\{-\log f(\mathbf{y})\} \quad (1)$$

where we assume that  $\mathbf{y}$  has a density  $f(\cdot)$ . Throughout the manuscript  $\log x = \log_2 x$  and the entropy is measured in bits. Assume that  $\dim A = m' \times n$  and  $\text{Rank} A = m$ . In some cases, this entropy is easily calculated or bounded:

1.  $A$  is an invertible matrix (i.e.,  $m' = m = n$ ). In this case the linear transformation just scales and shuffles  $\mathbf{x}$ , thus the entropy is only shifted,

$$h(A\mathbf{x}) = h(\mathbf{x}) + \log |A| \quad (2)$$

where  $|\cdot|$  denotes (absolute value of) determinant.

2.  $A$  does not have a full row-rank (i.e.,  $m' > m$ ). In this case there is a deterministic relation between the components of  $\mathbf{y}$  and thus

$$h(A\mathbf{x}) = -\infty \quad (3)$$

3.  $\mathbf{x} = \mathbf{x}^*$  is a Gaussian vector. The linear transformation  $A$  preserves the normality and so

$$h(A\mathbf{x}^*) = \frac{m}{2} \log(2\pi e |AR_x A^t|^{1/m}) \quad (4)$$

where  $R_x$  is the covariance matrix of  $\mathbf{x}^*$  and  $AR_x A^t$  is the covariance matrix of  $\mathbf{y}^* = A\mathbf{x}^*$ . Since for a given covariance, the Gaussian distribution maximizes the entropy, the expression

Manuscript received July 29, 1993; revised February 10, 1993. This work was supported in part by the Wolfson Research Awards administrated by the Israel Academy of Science and Humanities. This paper was presented in part at the International Symposium on Information Theory, San Antonio, TX, January 1993.

The authors are with the Department of Electrical Engineering—Systems, Tel-Aviv University, Tel-Aviv 69978, Israel.  
IEEE Log Number 9211273.

in (4) upper bounds the entropy of  $\mathbf{y} = A\mathbf{x}$  in the general case, i.e.,

$$h(A\mathbf{x}) \leq h(A\mathbf{x}^*) \quad (5)$$

where  $\mathbf{x}^*$  is now a Gaussian vector with the same covariance matrix as  $\mathbf{x}$ .

4. In the above three cases  $\mathbf{x}$  was an arbitrary random vector. In what follows we restrict  $\mathbf{x}$  to have independent components. If in addition  $y$  is scalar, i.e.,  $y = a_1 x_1 + \dots + a_n x_n$ , then the entropy-power inequality (EPI) can be used to lower bound its entropy. Specifically, by the EPI (see, e.g., [3], pp. 287),

$$P(y) \geq P(a_1 x_1) + \dots + P(a_n x_n) \quad (6)$$

where  $P(y) = \frac{1}{2\pi e} 2^{2h(y)}$  is the entropy-power of  $y$ . An equivalent form of the EPI [7] expresses (6) directly in terms of the entropy as

$$h(a^t \mathbf{x}) \geq h(a^t \tilde{\mathbf{x}}) \quad (7)$$

where  $\tilde{\mathbf{x}}$  is a Gaussian vector with independent components such that  $h(\tilde{x}_i) = h(x_i)$ ,  $i = 1 \dots n$  and  $a^t = (a_1, \dots, a_n)$ . An explicit calculation of the entropy in the RHS of (7) yields

$$h(a^t \tilde{\mathbf{x}}) = \frac{1}{2} \log 2\pi e (a^t P a) = \frac{1}{2} \log 2\pi e \left( \sum_{i=1}^n a_i^2 p_i \right) \quad (8)$$

where  $P$  is the covariance matrix of  $\tilde{\mathbf{x}}$ , i.e., it is a diagonal matrix whose  $i$ -th diagonal element is  $p_i = \frac{1}{2\pi e} 2^{2h(x_i)} = \text{Var}\{\tilde{x}_i\}$ , and  $h(x_i)$  is the entropy of  $x_i$ . The inequalities (6) and (7) become equalities iff  $\mathbf{x}$  is Gaussian.

We generalize the lower bound (7) to the case where  $\mathbf{y}$  may be a vector, and show below that  $h(A\mathbf{x}) \geq h(A\tilde{\mathbf{x}})$  for any  $A$ . Unlike what one may have expected, this inequality does not follow by just using in (7) the vector form of the EPI instead of the regular EPI. To see that, recall the vector form of the EPI (see e.g. [7])

$$h(\mathbf{u}_1 + \dots + \mathbf{u}_n) \geq h(\tilde{\mathbf{u}}_1 + \dots + \tilde{\mathbf{u}}_n) = \frac{m}{2} \log 2\pi e \left( \sum_{i=1}^n P(\mathbf{u}_i) \right) \quad (9)$$

where  $\mathbf{u}_i \in \mathcal{R}^m$ ,  $i = 1 \dots n$  are independent random vectors and  $\tilde{\mathbf{u}}_i \in \mathcal{R}^m$  are independent Gaussian vectors with (proportional) covariances  $R_i = P(\mathbf{u}_i) \cdot K$ , where  $K$  is any covariance matrix with a unity determinant (e.g.  $K = I$ ) and (the scalar)  $P(\mathbf{u}_i)$  is the entropy-power of the random vector  $\mathbf{u}_i$ ,

$$P(\mathbf{u}) = \frac{1}{2\pi e} 2^{\frac{2}{m} h(\mathbf{u})} \quad (10)$$

Now, let  $\mathbf{b}_1 \dots \mathbf{b}_n$  be the columns of  $A$ . Then, by the vector form of the EPI (9),

$$h(A\mathbf{x}) = h(x_1 \mathbf{b}_1 + x_2 \mathbf{b}_2 + \dots + x_n \mathbf{b}_n) \geq h(\tilde{x}_1 \mathbf{b}_1 + \tilde{x}_2 \mathbf{b}_2 + \dots + \tilde{x}_n \mathbf{b}_n) \quad (11)$$

At that point, one would like to proceed by replacing the RHS of (11) with  $h(\tilde{x}_1 \tilde{\mathbf{b}}_1 + \dots + \tilde{x}_n \tilde{\mathbf{b}}_n) = h(A\tilde{\mathbf{x}})$ . However, this transition fails since for  $m \geq 2$ ,  $h(x_i \mathbf{b}_i) = -\infty$ , or  $P(x_i \mathbf{b}_i) = 0$  (due to the deterministic relation between the components). Thus, a straightforward application of the vector form of the EPI leads to the trivial lower bound  $h(A\mathbf{x}) \geq -\infty$ .

Other simple attempts to get the desired generalization from the vector form of the EPI fail as well. Nevertheless, using a different approach, based on a double induction over the matrix dimensions, we prove the following.

*Theorem 1:* For any matrix  $A$ ,

$$h(A\bar{x}) \geq h(A\bar{x}) \quad (12)$$

The detailed proof is provided in Appendix A. Note that the RHS of (12) can be specified explicitly as  $h(A\bar{x}) = \frac{m}{2} \log(2\pi e |APA^t|^{1/m})$  where, as above,  $P$  is the covariance matrix of  $\bar{x}$  which is a diagonal matrix whose  $i$ -th diagonal element is  $p_i = \frac{1}{2\pi e} 2^{2h(x_i)}$ , and  $m = \text{Rank} A$ .

Equality in (12) holds in one of the following cases, which correspond to the three cases mentioned in the introduction:

1.  $\bar{x}$  is Gaussian ( $\bar{x} = \bar{x}$ ).
2.  $A$  is a non-singular square matrix (see (2)). More generally, we get equality in (12) if  $A$  contains all-zero columns, corresponding to components of  $\bar{x}$  that do not influence  $\underline{y}$ , but after these columns are removed  $A$  becomes a non-singular square matrix.
3.  $A$  does not have a full row-rank and so both sides of (12) equal  $-\infty$  (see (3)).

In the i.i.d. case  $P = p \cdot I$ , where  $p = \frac{1}{2\pi e} 2^{2h(x)}$  is the entropy-power of each component of  $\bar{x}$ , and so (12) becomes

$$\frac{1}{m} h(A\bar{x}) \geq h(x) + \frac{1}{2} \log(|AA^t|^{1/m}). \quad (13)$$

When  $|AA^t| = 1$ , e.g., in the case of orthonormal transformation, (13) is reduced to

$$\frac{1}{m} h(A\bar{x}) \geq h(x). \quad (14)$$

As the dimension of  $\bar{x}$  becomes large it may represent samples of a white stochastic process. In this case the matrix  $A$  represents a linear transformation of that process. When  $A$  represents a filtering operation, some projections of  $\bar{x}$  are transferred with unity gain and the rest are filtered away, and so  $|AA^t| = 1$ . Thus an interpretation of (14) is that after linear filtering the entropy (per degree-of-freedom) of a white process is increased.

The new inequality (12) leads, in general, to tighter bounds than the standard vector form EPI. Consider for example a vector  $\underline{z} = A\bar{x} + B\underline{y}$  where both  $\bar{x}, \underline{y}$  are independent vectors with  $n$  independent components, and  $A, B$  are nonsingular  $n \times n$  matrices. It is interesting to assess the value of  $h(\underline{z})$  for evaluating the capacity of some additive noise channels. In this case the standard EPI is applicable, leading to the bound

$$P(\underline{z}) \geq P(A\bar{x}) + P(B\underline{y}) = |AP_x A^t|^{1/n} + |BP_y B^t|^{1/n} \quad (15)$$

where  $P(\cdot)$  is the entropy power of a vector defined in (10) and  $P_x, P_y$  are diagonal matrices whose elements are the entropy powers of the components of  $\bar{x}$  and  $\underline{y}$ , respectively. Our generalization of the EPI leads to the bound

$$P(\underline{z}) \geq P(A\bar{x} + B\underline{y}) = |AP_x A^t + BP_y B^t|^{1/n}. \quad (16)$$

By the Minkowski inequality (see, e.g., [7, Theorem 5]), the bound (16) is tighter than (15), with equality iff  $AP_x A^t$  is proportional to  $BP_y B^t$ .

## II. APPLICATIONS TO LINEAR TRANSFORMATIONS OF A VECTOR WITH INDEPENDENT COMPONENTS

### A. Closeness to Normality after Transformation

It is well known that a Gaussian vector stays normal after linear transformations. It has also been observed that a non-Gaussian vector with independent components becomes "closer" to normality after passing through a linear transformation. The case of a non-Gaussian stochastic process whose samples are statistically independent (e.g.,

a non-Gaussian white noise) that passes through a linear system has drawn a special interest in the recent years in deconvolution problems. The closeness to normality of the output in this case has been characterized elegantly in [8], and has been used to derive techniques for deconvolving the effect of the linear system. In this section we use the generalization of the EPI to show that indeed a non-Gaussian vector with independent components becomes closer to normality, after a linear transformation, in a very specific sense where closeness is measured by the divergence (or "relative entropy", or "Kullback-Leibler distance") from Gaussianity.

We recall the definition of the divergence. Let  $\underline{y}$  be an  $n$ -dimensional random vector, and let  $\underline{y}^*$  be another vector. The divergence between these vectors is defined, (see, e.g., [5, pp. 231])

$$\mathcal{D}(\underline{y}; \underline{y}^*) \triangleq \int_{\mathcal{R}^n} f_{\underline{y}}(\alpha) \log \frac{f_{\underline{y}}(\alpha)}{f_{\underline{y}^*}(\alpha)} d\alpha \quad (17)$$

where  $f_{\underline{y}}(\cdot), f_{\underline{y}^*}(\cdot)$  are the corresponding probability density functions, and the divergence is measured in bits. For any two pdf's, the divergence is non-negative. The divergence from Gaussianity, i.e., the case where  $\underline{y}^*$  is Gaussian with the same first and second order moments as  $\underline{y}$ , can be expressed as

$$\mathcal{D}(\underline{y}; \underline{y}^*) = h(\underline{y}^*) - h(\underline{y}) \geq 0 \quad (18)$$

and it is zero iff  $\underline{y}$  is also Gaussian. If there is a deterministic linear dependency between the components of  $\underline{y}$  (e.g., when  $\underline{y}$  is the output of a system that does not have a full rank) then neither the integral in (17) nor the entropies in (18) are well defined, and the following more general definition of the divergence is used (see [10, pp. 20]):

$$\mathcal{D}(\underline{y}; \underline{y}^*) = E_{\underline{y}} \left\{ \log \frac{dF}{dF^*}(\underline{y}) \right\} \quad (19)$$

where  $F$  and  $F^*$  are the distributions of  $\underline{y}$  and  $\underline{y}^*$ ,  $\frac{dF}{dF^*}$  is the Radon-Nikodym derivative of the corresponding distributions, and the expectation is taken with respect to  $\underline{y}$ .

Using the generalization to the Entropy Power Inequality, derived in the previous section, we provide below an upper bound for the divergence from Gaussianity of a linear transformation of a vector  $\underline{x} = x_1 \dots x_n$  with independent components. In stating this result we denote by  $\underline{x}^*$  a Gaussian vector with independent components, such that  $\text{Var}\{x_i\} = \text{Var}\{x_i^*\}$ . Unlike the previous lower bound for the entropy, this upper bound is not trivial even when the transformation does not have a full rank.

*Theorem 2:* For any matrix  $A$ ,

$$\frac{1}{m} \mathcal{D}(A\bar{x}; A\bar{x}^*) \leq \frac{1}{2} \log \left( \frac{|A_o R_x A_o^t|^{1/m}}{|A_o P_x A_o^t|^{1/m}} \right) \leq \max_{i=1, \dots, n} \mathcal{D}(x_i; x_i^*) \quad (20)$$

where  $m = \text{Rank} A$ ,  $A_o$  is any  $m \times n$  matrix whose rows span the row space of  $A$ ,  $P_x$  is a diagonal matrix whose diagonal elements are  $\{p_i\}$  the entropy powers of the components of  $\bar{x}$ , and  $R_x$  is the diagonal covariance matrix of  $\bar{x}^*$  whose diagonal elements are  $\{\sigma_i^2\}$ , the powers of the components of  $\bar{x}$ .

Note that if the components of  $\bar{x}$  are i.i.d., (20) is reduced to

$$\frac{1}{m} \mathcal{D}(A\bar{x}; A\bar{x}^*) \leq \mathcal{D}(x; x^*) \quad (21)$$

where  $x$  is any component and equality holds if  $x$  is Gaussian ( $\mathcal{D}(x; x^*) = 0$ ) or if  $A$  is invertible (after all its zero columns, if any, are removed). This theorem follows straight-forwardly from Theorem 1, and its detailed proof is given in Appendix B.

Theorem 2 can be used to show that an i.i.d. process becomes closer to normality, in information divergence sense, after passing through a linear-time-invariant system. For this we consider the limit,

as  $n$  goes to infinity, of the normalized divergence per degree-of-freedom of  $n$  samples of the output process. The inequality (21) is satisfied by the normalized divergence for any  $n$  and so it is satisfied in the limit. The interpretation of inequality (21) in this case is that a white process becomes "more Gaussian" after filtering, in the sense that its normalized divergence from Gaussianity, per degree-of-freedom, decreases. Note that if the filter is invertible, the normalized divergence of the entire output process does not change. Yet, the divergence from Gaussianity of a finite number of samples becomes smaller, since these samples are obtained from the entire input process by a non-invertible transformation.

Finally, it is interesting to note that Theorem 2 yields a stronger result than a straight-forward application of the data processing theorem for the divergence. For example, when  $\underline{x}$  is i.i.d., the data processing theorem for the divergence implies

$$\mathcal{D}(A\underline{x}; A\underline{x}^*) \leq \mathcal{D}(\underline{x}; \underline{x}^*) = n \cdot \mathcal{D}(x; x^*). \quad (22)$$

Since  $n \geq m = \text{Rank} A$ , the bound (21) is tighter.

### B. Mutual-Information between Orthogonal Projections of an Independent Vector

A pair of orthogonal projections of uncorrelated Gaussian vector are independent and therefore the mutual-information between them is zero. This may not be true, however, for non-Gaussian noise. In this section we show that the projection of a non-Gaussian vector with independent components into two subspaces that span the entire space, results in two vectors whose mutual information is lower bounded away from zero. Note that since the mutual-information is invariant to the representation, it is only a function of the pair of linear sub-spaces spanned by the projections.

Let  $x$  be a random variable, and let  $\underline{x}$  be an  $n$ -dimensional vector of i.i.d. samples, distributed as  $x$ . Let  $A_l$  and  $A_h$  be two matrices, each with  $n$  columns, where  $\text{Rank} A_l = r$  ( $r < n$ ),  $\text{Rank} A_h = n - r$ , and the space spanned by the rows of  $A_l$  is orthogonal to the space spanned by the rows of  $A_h$ . The rows of  $A_l$  and  $A_h$  thus span the entire space. The projections are denoted  $\underline{y}_l = A_l \underline{x}$  and  $\underline{y}_h = A_h \underline{x}$ .

One motivation to consider the mutual information  $I(\underline{y}_l; \underline{y}_h)$  comes from the following example. Let  $\underline{X} = [X_0, \dots, X_{n-1}]^t$  be the DFT of  $\underline{x} = [x_0, \dots, x_{n-1}]^t$  i.e.

$$X_k = \frac{1}{\sqrt{n}} \sum_{m=0}^{n-1} x_m e^{-j \frac{2\pi}{n} km}, \quad k = 0, \dots, n-1,$$

where  $j = \sqrt{-1}$ . The random vector  $\underline{X}$  represents the spectral content of the vector  $\underline{x}$ . In general, it is interesting to find the mutual information between mutually exclusive spectral components of the i.i.d. vector  $\underline{x}$ . For example, the mutual information  $I(X_0; X_1, \dots, X_{n-1})$ , i.e. the mutual information between the DC-component and the rest of the spectral components, has been considered in [12].

Define the divergence from Gaussianity of  $\underline{y}_l$  (normalized to bits-per-sample) as

$$\mathcal{D}_l = \frac{1}{r} \mathcal{D}(\underline{y}_l; \underline{y}_l^*) = \frac{1}{r} \mathcal{D}(A_l \underline{x}; A_l \underline{x}^*). \quad (23)$$

A similar definition can be made for  $\mathcal{D}_h$ . Now, in some applications  $r$  is fixed, while  $n$  becomes large, and so  $\mathcal{D}_l$  can be made arbitrarily small. For example, fix  $r = 1$  and let  $A_l$  be the DC-component, i.e.,  $y_l = X_0 = \frac{1}{\sqrt{n}} \sum_{i=1}^n x_i$ . Then by the strong form of the central limit theorem of [1],  $\mathcal{D}_l \rightarrow 0$  as  $n \rightarrow \infty$ . A projection  $A_l$  for which  $\mathcal{D}_l \rightarrow 0$  as  $n \rightarrow \infty$ , is referred to as "asymptotically Gaussian projection."

The following theorem underbounds the mutual-information between  $\underline{y}_l$  and  $\underline{y}_h$ , per degree-of-freedom (dimension) of  $\underline{y}_l$ .

*Theorem 3:*

$$\frac{1}{r} I(A_l \underline{x}; A_h \underline{x}) \geq \mathcal{D}(x; x^*) - \mathcal{D}_l. \quad (24)$$

The theorem is proved in Appendix C. Note that by Theorem 2 the RHS of (24) is positive, bounded away from zero, unless  $x$  is Gaussian. Also, if  $A_l$  is an asymptotically Gaussian projection, the lower bound becomes the divergence from Gaussianity of  $x$ .

Returning to the example that motivated this problem, we have calculated explicitly the mutual information between the DC-component and the rest of the spectral components for a uniformly distributed i.i.d. vector. When the vector dimension  $n = 2$ ,  $I(X_0; X_1) = I(x_0 + x_1; x_0 - x_1) = \log(\frac{e}{2}) \cong 0.44$  bit. For dimension  $n = 3$  the mutual information is computed numerically, using the relation

$$I(X_0; X_1, X_2) = I\left(x_0 + x_1 + x_2; x_0 - \frac{x_0 + x_1 + x_2}{3}, x_1 - \frac{x_0 + x_1 + x_2}{3}\right) \cong 0.6 \text{ bit.}$$

In both cases the mutual information is greater than  $\mathcal{D}(x; x^*) = 0.254$ , the divergence between a uniform distribution and a Gaussian distribution having the same variance.

Notice that Theorem 3 above provides a lower bound on the mutual information, whose main properties are that it is greater than zero, and it depends on the divergence from Gaussianity of the distribution of each sample, and on the dimension of  $A_l$ , but it does not depend explicitly on the projections themselves. However, the general problem of estimating the mutual-information between orthogonal projections of a white vector (or process) is still open, especially, since from the example above, the lower bound seems to be untight. A somewhat related subject is to find the mutual-information between a subset and its complement in a given set of elements, treated in [7].

### III. A GENERALIZATION OF THE FISHER-INFORMATION-INEQUALITY

The duality between the EPI and various information inequalities has been pointed out in [7]. One example of such dual inequality is the Fisher-Information-Inequality (FII)

$$K(\underline{X} + \underline{Y})^{-1} \geq K(\underline{X})^{-1} + K(\underline{Y})^{-1} \quad (25)$$

where  $\underline{X}$  and  $\underline{Y}$  are independent random vectors, and  $K$  is the  $n \times n$  dimensional Fisher-information-matrix of an  $n$ -dimensional random vector having a differentiable density  $f$ , with respect to a translation parameter, defined as

$$K = E \left\{ \frac{1}{f^2} \nabla f \cdot \nabla f^t \right\} \quad (26)$$

where  $\nabla f$  is the  $n$ -dimensional gradient vector of  $f$  (see [7]). The scalar Fisher-information is defined as  $J = \text{tr}\{K\} = E \left\{ \frac{1}{f^2} \|\nabla f\|^2 \right\}$ . The FII (25), whose proof is relatively simple, is actually used to prove the EPI (see [11] and [2]).

The generalized EPI and this duality motivated us to show the following generalization of the FII.

*Theorem 4:* Let  $\underline{x} = x_1 \dots x_n$  be a vector with independent components having a (diagonal) Fisher information matrix  $K$ . Then, for any matrix  $A$

$$K(A\underline{x})^{-1} \geq K(A\underline{\hat{x}})^{-1} = AK(\underline{x})^{-1}A^t \quad (27)$$

where  $\underline{\hat{x}} = \hat{x}_1 \dots \hat{x}_n$  is a Gaussian vector with independent components, such that  $J(\hat{x}_i) = \text{Var}\{\hat{x}_i\}^{-1} = J(x_i)$ ,  $i = 1 \dots n$ .

Note that the matrix inequality (27) is in the sense that the difference matrix is positive semi-definite. The detailed proof of this theorem is given in [13], and here we sketch its structure. Similarly to the derivation in [11], [2] and [6], where the basic FII was shown, we show that

$$\underline{b}_i^t \cdot \frac{\nabla f(\underline{y})}{f(\underline{y})} = E \left\{ \frac{f'(x_i)}{f(x_i)} \middle| \underline{y} \right\}, \quad \text{for } i = 1 \dots n \quad (28)$$

where  $\underline{b}_i$  is the  $i$ -th column of  $A$ , and the conditional expectation is over  $x_i$  given  $\underline{y} = A\underline{x}$ . This equality can be written in a matrix form as

$$A^t \cdot \frac{\nabla f(\underline{y})}{f(\underline{y})} = E \left\{ \frac{\nabla f(\underline{x})}{f(\underline{x})} \middle| \underline{y} \right\}. \quad (29)$$

Using (29) and the Cauchy-Schwarz inequality

$$\begin{aligned} A^t \left( \frac{\nabla f(\underline{y})}{f(\underline{y})} \right) \left( \frac{\nabla f(\underline{y})}{f(\underline{y})} \right)^t A &= E \left\{ \frac{\nabla f(\underline{x})}{f(\underline{x})} \middle| \underline{y} \right\} E \left\{ \frac{\nabla f(\underline{x})^t}{f(\underline{x})} \middle| \underline{y} \right\} \\ &\leq E \left\{ \left( \frac{\nabla f(\underline{x})}{f(\underline{x})} \right) \left( \frac{\nabla f(\underline{x})}{f(\underline{x})} \right)^t \middle| \underline{y} \right\}. \end{aligned} \quad (30)$$

Averaging (30) over  $\underline{y}$  gives

$$A^t K(\underline{y}) A \leq K(\underline{x}). \quad (31)$$

Finally, multiplying (31) from the left by  $(AK(\underline{x})^{-1}A^t)^{-1}AK(\underline{x})^{-1}$ , multiplying from the right by  $K(\underline{x})^{-1}A^t(AK(\underline{x})^{-1}A^t)^{-1}$ , and taking the inverse we get (27). Note that the Fisher information matrix of the Gaussian vector  $A\underline{x}$  in (27) is given directly by its inverse covariance matrix.

As in Theorem 1, equality in (27) holds if  $\underline{x}$  is Gaussian or if  $A$  is invertible. Note that in the i.i.d. case  $K = J(\underline{x}) \cdot I$ , and if we further assume that  $A$  is orthonormal (i.e.,  $AA^t = I$ ), we can rewrite inequality (27) in a scalar form as  $\frac{1}{m}J(A\underline{x}) \leq J(\underline{x})$ .

As in the standard EPI, one may hope to prove the generalized EPI, i.e., to prove Theorem 1 from the generalized FII. Recently, indeed, we were able to do that, and through the derivation we proved additional generalizations of information theoretic inequalities. This alternative proof of Theorem 1 is provided in [14].

#### APPENDIX A: PROOF OF THEOREM 1

We prove (12) for a matrix  $A$ , whose number of rows is  $m = \text{Rank } A$ . The case where the number of rows  $m' > \text{Rank } A$ , i.e.,  $A$  does not have a full row-rank, is trivial since both sides of (12) are  $-\infty$  (see (3)).

The proof is by *double induction* over  $m$  and  $n$ . The *induction boundary conditions* are the line  $m = 1$  (any  $n$ ) and the line  $m = n$  in the plain  $(m, n) \in \mathcal{N}^2$ . In the case  $m = 1$  the inequality holds by the regular EPI since  $A$  is a row matrix. In the case  $m = n$  the matrix is invertible and so (12) holds with equality. We show below that if (12) holds for any  $(m-1) \times (n-1)$  and  $m \times (n-1)$  matrices, then it also holds for any  $m \times n$  matrix. This is the *induction step*. Fig. 1 shows a path in the plain  $\mathcal{N}^2$  from the boundary lines to an arbitrary point  $(m, n)$ , which is followed by the induction steps to prove the theorem for any  $m \times n$  matrix. Since  $m$  and  $n$  are arbitrary, the theorem holds for any matrix, provided that the induction step is proved.

To prove the induction step, some matrix manipulations used in Gaussian elimination, are needed. Denote by  $\{a_{i,j}\}$ ,  $i = 1 \dots m$ ,  $j = 1 \dots n$ , the elements of the matrix  $A$ , and let  $\text{Rank } A = m \geq 2$ . Suppose the last column of  $A$  is not zero. Otherwise, i.e., if  $a_{i,n} = 0$  for all  $i$ , then  $\underline{y}$  does not depend on  $x_n$  and  $A$  is actually  $m \times (n-1)$  matrix for which the inequality holds by the induction assumption. Now if  $a_{m,n} = 0$ , permute a pair of rows of  $A$  and the pair of

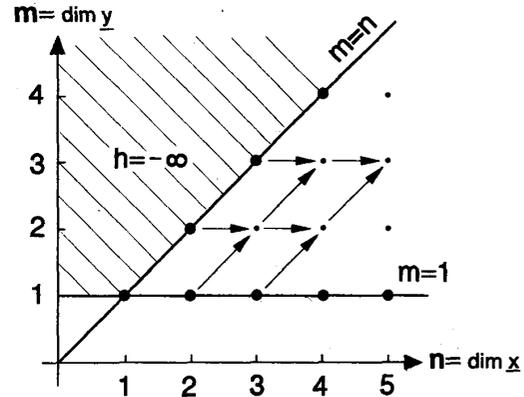


Fig. 1. Path from boundary lines followed by induction steps.

corresponding components of  $\underline{y}$ , so that after permutation  $a_{m,n} \neq 0$ . This permutation, if needed, does not affect the entropy.

The next step is to use row operations and make the first  $(m-1)$  elements of the last column to be zero. This is possible since we assured above that  $a_{m,n} \neq 0$ . Denote by  $\hat{A} = TA$  the matrix after the row operations, where the matrix  $T$  has the form

$$T = \begin{pmatrix} 1 & & & \alpha_1 \\ & 1 & 0 & \alpha_2 \\ & & \ddots & \vdots \\ 0 & & & 1 \end{pmatrix}, \quad |T| = 1 \quad (32)$$

Observe that since  $|T| = 1$  the row operations do not change the entropy,

$$h(A\underline{x}) = h(T^{-1}\hat{A}\underline{x}) = h(\hat{A}\underline{x}) - \log |T| = h(\hat{A}\underline{x}). \quad (33)$$

Define some submatrices of  $\hat{A}$ , as follows

$$\hat{A} = \left( \begin{array}{c|c} A^- & \begin{matrix} 0 \\ \vdots \\ 0 \end{matrix} \\ \hline \hat{\underline{a}}_m^t & a_{m,n} \end{array} \right) = \left( \begin{array}{c|c} B & \hat{\underline{b}}_n \end{array} \right)$$

where  $\hat{\underline{a}}_m = (a_{m,1}, \dots, a_{m,n-1})^t$  is the last row of  $A$  without the last term,  $\hat{\underline{b}}_n = (0, 0, \dots, 0, a_{m,n})^t$  is the last column of  $\hat{A}$ ,  $B$  is obtained by dropping the last column of  $\hat{A}$  ( $\dim B = m \times (n-1)$ ),  $A^-$  is obtained by dropping the last row of  $B$  ( $\dim A^- = (m-1) \times (n-1)$ ) and  $\underline{x}^- = x_1, \dots, x_{n-1}$  is the vector  $\underline{x}$  without the last component. Now the matrices  $\hat{A}$  and  $A^-$  have a full row-rank since they are obtained by row operations from the matrix  $A$ , which has a full row-rank. The matrix  $B$ , however, may either have a full row-rank, if its last row,  $\hat{\underline{a}}_m$ , does not depend linearly on the other rows (i.e., on  $A^-$ ), or a deficient rank, if its last row linearly depends on the other rows.

Note that all the components of  $\hat{A}\underline{x}$ , with the exception of the last one, are independent of  $x_n$ . Also observe that, by the induction assumption, both the matrix  $A^-$  (of size  $(m-1) \times (n-1)$ ) and the matrix  $B$  (of size  $m \times (n-1)$ ) satisfy (12), i.e.,  $h(A^-\underline{x}^-) \geq h(A^-\hat{\underline{x}}^-)$  and  $h(B\underline{x}^-) \geq h(B\hat{\underline{x}}^-)$ .

To utilize the induction assumptions, we need to express the entropy  $\hat{A}\underline{x}$  in terms of entropies associated with lower dimensional

matrices, e.g. the entropy of  $A^-x^-$ . Using the chain rule,

$$h(\hat{A}\underline{x}) = h(\hat{y}_1, \dots, \hat{y}_m) = h(\hat{y}_1, \dots, \hat{y}_{m-1}) + h(\hat{y}_m | \hat{y}_1, \dots, \hat{y}_{m-1}) \quad (34)$$

and since  $\hat{y}_m = \underline{a}_m^t \underline{x} = \hat{a}_m^t \underline{x}^- + a_{m,n} x_n$  and  $(\hat{y}_1, \dots, \hat{y}_{m-1}) = A^-x^-$ , we can rewrite (34) as

$$h(\hat{A}\underline{x}) = h(A^-x^-) + h(\hat{a}_m^t \underline{x}^- + a_{m,n} x_n | A^-x^-). \quad (35)$$

Notice that  $a_{m,n} x_n$  in the RHS of (35) is independent of both  $\hat{a}_m^t \underline{x}^-$  and the condition  $A^-x^-$ .

Suppose first that the last row of the matrix  $B$  linearly depends on the other rows. In this case the term  $\hat{a}_m^t \underline{x}^-$  in (37) linearly depends on  $A^-x^-$  and does not affect the entropy. Thus,

$$h(\hat{A}\underline{x}) = h(A^-x^-) + h(a_{m,n} x_n) = h(A^-x^-) + h(x_n) + \log |a_{m,n}|. \quad (36)$$

Utilizing the induction assumption, asserting  $h(A^-x^-) \geq h(A^-x^-)$ , and by (33)

$$h(A\underline{x}) \geq h(A^-x^-) + h(x_n) + \log |a_{m,n}| = h(A\underline{x}) \quad (37)$$

where the second equality follows by applying (36) to  $h(A\underline{x})$  and since  $h(x_n) = h(\bar{x}_n)$ . The induction step for this case is proved.

Consider now the second case where  $B$  has a full row-rank. Proceeding from (35), we use a conditional version of the EPI (originally presented in [2], see also [3] pp. 289) to lower bound the entropy of the sum of independent terms in the RHS of (35)

$$h(\hat{A}\underline{x}) \geq h(A^-x^-) + \frac{1}{2} \log \left( 2^{2h(\hat{a}_m^t \underline{x}^- | A^-x^-)} + 2^{2h(a_{m,n} x_n)} \right). \quad (38)$$

Since  $B\underline{x}^-$  is a concatenation of  $A^-x^-$  and  $\hat{a}_m^t \underline{x}^-$ , we can use again the chain rule to get

$$h(\hat{A}\underline{x}) \geq h(A^-x^-) + \frac{1}{2} \log \left( 2^{2[h(B\underline{x}^-) - h(A^-x^-)]} + a_{m,n}^2 2^{2h(x_n)} \right). \quad (39)$$

The RHS of (39) is clearly monotonically increasing with  $h(B\underline{x}^-)$ . Similarly, the function  $\alpha(t) = t + \frac{1}{a} \log(b2^{-at} + c)$ ,  $a, b, c > 0$ , has a positive derivative for all  $t$ , and so the RHS of (39) is also monotonically increasing with  $h(A^-x^-)$ . Since by the induction assumption  $h(A^-x^-) \geq h(A^-x^-)$  and  $h(B\underline{x}^-) \geq h(B\underline{x}^-)$ , we can lower bound (39)

$$h(A\underline{x}) \geq h(A^-x^-) + \frac{1}{2} \log \left( 2^{2[h(B\underline{x}^-) - h(A^-x^-)]} + a_{m,n}^2 2^{2h(x_n)} \right). \quad (40)$$

To complete the induction step, observe that the conditional version of the EPI used in the transition from (35) to (38) holds with equality for the Gaussian vector  $\underline{x}^-$  and thus the RHS of (40) is  $h(A\underline{x})$ , as desired.  $\square$

#### APPENDIX B: PROOF OF THEOREM 2

Assume, first, that  $A$  has a full row-rank,  $\dim A = m \times n$ . By Theorem 1,

$$h(A\underline{x}) \geq h(A\underline{x}) = \frac{m}{2} \log(2\pi e |APA^t|^{1/m}) \quad (41)$$

where  $P$  is the diagonal covariance matrix of  $\underline{x}$ , whose diagonal elements are  $p_1 \dots p_n$ . Thus,

$$h(A\underline{x}^*) - h(A\underline{x}) \leq h(A\underline{x}^*) - h(A\underline{x}) = \frac{m}{2} \log \left( \frac{|AR_x A^t|^{1/m}}{|APA^t|^{1/m}} \right). \quad (42)$$

where  $h(A\underline{x}^*) = \frac{m}{2} \log(2\pi e |AR_x A^t|^{1/m})$  and  $R_x$  is the diagonal covariance matrix of  $\underline{x}^*$  whose diagonal elements are  $\sigma_1^2 \dots \sigma_n^2$  (the

powers of the components of  $\underline{x}$ ). Using the identity (18) and the fact that  $(A\underline{x})^* = A\underline{x}^*$ , we get the first inequality in (20)

$$\frac{1}{m} \mathcal{D}(A\underline{x}; A\underline{x}^*) \leq \frac{1}{2} \log \left( \frac{|AR_x A^t|^{1/m}}{|APA^t|^{1/m}} \right). \quad (43)$$

Now, if the components of  $\underline{x}$  are i.i.d., then  $R_x = \sigma^2 \cdot I$ ,  $P = p \cdot I$  and so,

$$\frac{1}{m} \mathcal{D}(A\underline{x}; A\underline{x}^*) \leq \frac{1}{2} \log \left( \frac{\sigma^2}{p} \right) = \mathcal{D}(x; x^*), \quad (44)$$

which proves (21).

For the general case, as shown in Lemma 1 below,

$$\frac{|AR_x A^t|^{1/m}}{|APA^t|^{1/m}} \leq \max_{i=1 \dots n} \frac{\sigma_i^2}{p_i} \quad (45)$$

and the second inequality in (20) follows since  $\mathcal{D}(x; x_i^*) = \frac{1}{2} \log \left( \frac{\sigma_i^2}{p_i} \right)$ . Note that while the second inequality in (20) is less tight than the first, it is independent of the transformation  $A$ .

Consider now the case where  $A$  does not have a full row rank, i.e.,  $\text{Rank } A = m$  is less than the number of rows. In this case there is an invertible transformation  $T$ , such that  $A = TA_o$ , where  $A_o$  is an  $m \times n$  matrix with full row rank. By the definition (19), the divergence is invariant to invertible transformations, and so,

$$\mathcal{D}(\underline{y}; \underline{y}^*) = \mathcal{D}(T^{-1}\underline{y}; T^{-1}\underline{y}^*) = \mathcal{D}(A_o \underline{x}; A_o \underline{x}^*). \quad (46)$$

Since  $A_o$  has a full row rank, we can apply the derivation above to  $\mathcal{D}(A_o \underline{x}; A_o \underline{x}^*)$ , and prove the Theorem.  $\square$

In the proof we have used the following lemma.

**Lemma 1:** Let  $\Lambda$  and  $P$  be  $n \times n$  positive, diagonal matrices, with diagonal elements  $\lambda_1 \dots \lambda_n$  and  $p_1 \dots p_n$  respectively,  $\lambda_i, p_i > 0 \forall i$ . Then for any  $m \times n$  matrix  $A$ ,

$$\frac{|A\Lambda A^t|^{1/m}}{|APA^t|^{1/m}} \leq \max_{i=1 \dots n} \frac{\lambda_i}{p_i} \quad (47)$$

*Proof:* Define

$$r_m \triangleq \max_{i=1 \dots n} \frac{\lambda_i}{p_i}. \quad (48)$$

Clearly,  $r_m \cdot p_i - \lambda_i \geq 0$  for any  $i = 1 \dots n$ , and so the matrix  $r_m \cdot P - \Lambda$  is non-negative definite. As a result, the matrix  $A(r_m \cdot P - \Lambda)A^t$  is non-negative definite for any choice of an  $m \times n$  matrix  $A$ . Thus, we may write the matrix inequality

$$A(r_m \cdot P - \Lambda)A^t \geq 0 \implies 0 \leq A\Lambda A^t \leq r_m \cdot APA^t. \quad (49)$$

The inequality (49) implies a similar inequality for determinants (since  $|K_1 + K_2|$  is greater or equal both  $|K_1|$  and  $|K_2|$ ,  $K_1, K_2$  semi-definite matrices)

$$|A\Lambda A^t| \leq |r_m APA^t| = (r_m)^m |APA^t| \quad (50)$$

and (47) is proved.

#### APPENDIX C: PROOF OF THEOREM 3

Using the decomposition of the mutual information to entropies and by (18), one can express the mutual-information  $I(\underline{y}_i; \underline{y}_h)$  in terms of divergence as:

$$I(\underline{y}_i; \underline{y}_h) = I(\underline{y}_i^*; \underline{y}_h^*) - \mathcal{D}(\underline{y}_i; \underline{y}_i^*) - \mathcal{D}(\underline{y}_h; \underline{y}_h^*) + \mathcal{D}(\underline{y}_i; \underline{y}_h; \underline{y}_i^*, \underline{y}_h^*). \quad (51)$$

Examine now each term in the RHS of (51). Since orthogonality implies independence for zero-mean Gaussian vectors,

$$I(\underline{y}_i^*; \underline{y}_h^*) = 0. \quad (52)$$

From (23),  $\mathcal{D}(\underline{y}_l; \underline{y}_l^*) = r\mathcal{D}_l$ . By applying theorem 2 to  $A_h$ ,

$$\mathcal{D}(\underline{y}_h; \underline{y}_h^*) \leq (n-r)\mathcal{D}(x; x^*). \quad (53)$$

Finally,

$$\mathcal{D}(\underline{y}_l, \underline{y}_h; \underline{y}_l^*, \underline{y}_h^*) = \mathcal{D}(\underline{y}; \underline{y}^*) = \mathcal{D}(\underline{x}; \underline{x}^*) = n\mathcal{D}(x; x^*) \quad (54)$$

since  $A_l$  and  $A_h$  compose together an invertible transformation which preserves the divergence. Combining (23) and (51)-(54) yields the desired result.  $\square$

#### ACKNOWLEDGMENT

We thank S. Shamai (Shitz) and the anonymous referee for pointing out the example in the end of Section I. A discussion with U. Shaked regarding the proof of Lemma 1 is also acknowledged.

#### REFERENCES

- [1] A.R. Barron, "Entropy and the central limit theorem," *Ann. Probab.*, vol. 14, no. 1, pp. 336-342, 1986.
- [2] N. M. Blachman, "The convolution inequality for entropy powers," *IEEE Trans. Inform. Theory*, vol. IT-11, pp. 267-271, 1965.
- [3] R. E. Blahut, *Principles and Practice of Information Theory*. Reading, MA: Addison Wesley, 1987.
- [4] J. H. Conway and N. J. A. Sloane, "Voronoi regions of lattices, second moments of polytopes, and quantization," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 211-226, Mar. 1982.
- [5] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
- [6] A. Dembo, "Simple proof of the concavity of the entropy power with respect to added Gaussian noise," *IEEE Trans. Inform. Theory*, vol. 35, pp. 887-888, 1989.
- [7] T. M. Cover and J. A. Thomas, "Information theoretic inequalities," *IEEE Trans. Inform. Theory*, vol. 37, pp. 1501-1518, Nov. 1991.
- [8] D. Donoho, "On minimum entropy deconvolution," in *Applied Time Series Analysis II*. New York: Academic, 1981, pp. 565-608.
- [9] A. Gersho, "Asymptotically optimal block quantization," *IEEE Trans. Inform. Theory*, vol. IT-25, pp. 373-380, July 1979.
- [10] M. S. Pinsker *Information and Information Stability of Random Variables*. San Francisco, CA: Holden-Day, 1964.
- [11] A. J. Stam, "Some inequalities satisfied by the quantities of information of Fisher and Shannon," *Inform. Contr.*, vol. 2, pp. 101-112, June 1959.
- [12] R. Zamir and M. Feder, "Rate distortion performance in coding band-limited sources by sampling and dithered quantization," *IEEE Trans. Inform. Theory*, to be published.
- [13] —, "A generalization of the Fisher information inequality," Dep. EE-Syst., Tel-Aviv Univ., Tech. Rep., EE-S-93-03, Jan. 1993.
- [14] R. Zamir and M. Feder, "A generalization of information theoretic inequalities to linear transformation of independent vectors," in *Proc. Sixth Joint Swedish-Russian Int. Workshop on Inform. Theory*, Mölle, Sweden, Aug. 22-27, 1993, pp. 254-258.

## An Asymptotic Expression for the Information and Capacity of a Multidimensional Channel with Weak Input Signals

Vyacheslav V. Prelov and Edward C. van der Meulen, *Fellow, IEEE*

**Abstract**—An asymptotic expression is derived for the Shannon mutual information between the input and output signals for a relatively large class of continuous alphabet memoryless channels in the case of weak input signals, when the input space is multidimensional. We extend hereby a result of Ibragimov and Khas'minskii from the one-dimensional to the  $N$ -dimensional case. The asymptotic expression obtained relates the Shannon mutual information function and the Fisher information matrix. This expression is used to derive an asymptotic expression for the capacity of continuous alphabet memoryless channels with vector-valued weak input signals. This asymptotic capacity involves the largest eigenvalue of the Fisher information matrix evaluated at the zero input signal.

**Index Terms**—Continuous alphabet memoryless channel, weak multidimensional input signal, mutual information, Fisher information, mean energy constraint, peak power constraint, asymptotic capacity.

#### I. INTRODUCTION

The investigation of the asymptotic behavior of the capacity of communication channels for cases in which certain parameters characterizing the transmission can be designated as small has been of interest for some time. The reason for this is that the capacity of continuous alphabet memoryless channels can be calculated explicitly only in a relatively small number of cases.

When the input signal satisfies certain constraints, the evaluation of capacity necessitates the optimization over all probability distributions from a certain class. This is the reason why for most channels the capacity cannot be explicitly calculated. However, as is well-known, there is one case for which the capacity can be calculated exactly, namely for a channel with additive Gaussian noise and an energy constrained input [7].

The first studies regarding the asymptotic capacity of channels satisfying certain input constraints when the signal is weak (or alternatively when the noise is large) were carried out by Prelov [5].

Probably the most interesting from a theoretical and practical point of view is the investigation of the capacity of a channel with very weak input signals, because in the opposite extreme case of a channel with very strong input signals (or alternatively with very low noise) reliable transmission can be achieved even without the use of information-theoretic methods.

In this connection it should be mentioned that the asymptotic behavior of some characteristics of certain kinds of very noisy channels was investigated in Viterbi and Omura [10, Chapter 3] and by Viterbi [11]. However, the approach taken by those authors is different from ours.

Manuscript received July 23, 1991; revised February 4, 1993. This work was supported in part by Project OT/89/12, Onderzoeksfonds K.U. Leuven. This paper was presented in part at the IEEE Information Theory Workshop, Salvador, Brazil, June 21-26, 1992, and at the Benelux 14th Symposium on Information Theory, Veldhoven, The Netherlands, May 17-18, 1993, and at the 6th Joint Swedish-Russian International Workshop on Information Theory, Mölle, Sweden, August 22-27, 1993.

V. V. Prelov is with the Institute for Problems of Information Transmission, Russian Academy of Sciences, Moscow GSP-4, Russia.

E. C. van der Meulen is with the Department of Mathematics, Katholieke Universiteit Leuven, Celestijnenlaan 200B, B-3001 Leuven, Belgium.

IEEE Log Number 9211434.