A Generalization of the Entropy Power Inequality with Applications*

Ram Zamir and Meir Feder
Department of Electrical Engineering - Systems
Tel-Aviv University
Tel-Aviv, 69978, ISRAEL

Abstract

We prove the following generalization of the Entropy Power Inequality:

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

where $h(\cdot)$ denotes (joint-) differential-entropy, $\underline{x} = x_1 \ldots x_n$ is a random vector with independent components, $\underline{\tilde{x}} = \tilde{x}_1 \ldots \tilde{x}_n$ is a Gaussian vector with independent components such that $h(\tilde{x}_i) = h(x_i)$, $i = 1 \ldots n$, 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 non overlapping spectral components of a non-Gaussian white process. Finally, we describe a dual generalization of the Fisher Information Inequality.

Key Words: Entropy Power Inequality, Non-Gaussianity, Divergence, Fisher Information Inequality.

*This research was supported in part by the Wolfson Research Awards administrated by the Israel Academy of Science and Humanities, at Tel-Aviv University. This work was partially presented at the International Symposium on Information Theory, San Antonio TX., January 1993

1 The Generalization of the Entropy Power Inequality

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

$$h(\underline{y}) \triangleq E\{-\log f(\underline{y})\} \tag{1}$$

where we assume that $\underline{y}$ 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 $\mathrm{Rank} A = m$. 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 lineartransformation just scales and shuffles $\underline{x}$, thus the entropy is only shifted,

$$h(A\underline{x}) = h(\underline{x}) + \log |A| \tag{2}$$

   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 $\underline{y}$ and thus

$$h(A\underline{x}) = -\infty \quad . \tag{3}$$

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

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

   where R_x is the covariance matrix of $\underline{x}^*$ and AR_xA^t is the covariance matrix of $\underline{y}^* = A\underline{x}^*$. Since for a given covariance, the Gaussian distribution maximizes the entropy, the expression in (4) upper bounds the entropy of $\underline{y} = A\underline{x}$ in the general case, i.e.,

$$h(A\underline{x}) \le h(A\underline{x}^*) \tag{5}$$

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

4. In the above three cases $\underline{x}$ was an arbitrary random vector. In what follows we restrict $\underline{x}$ to have independent components. If in addition y is scalar, i.e., $y = a_1x_1 + \ldots + a_nx_n$, then the

entropy-power inequality (EPI) can be used to lower bound its entropy. Specifically, by the EPI (see e.g. [1], pp. 287),

$$P(y) \geq P(a_1 x_1) + \ldots + P(a_n x_n) \tag{6}$$

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

$$h(\underline{a}^t \underline{x}) \geq h(\underline{a}^t \underline{\tilde{x}}) \tag{7}$$

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

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

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

We generalize the lower bound (7) to the case where $\underline{y}$ may be a vector, and show below that $h(A\underline{x}) \geq h(A\underline{\tilde{x}})$ 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. [2])

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

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

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

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

$$h(A\underline{x}) = h(x_1\underline{b}_1 + x_2\underline{b}_2 + \ldots + x_n\underline{b}_n) \geq h\left(\widetilde{x_1\underline{b}_1} + \widetilde{x_2\underline{b}_2} + \ldots + \widetilde{x_n\underline{b}_n}\right) \tag{11}$$

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

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:

Theorem 1 *For any matrix A,*

$$h(A\underline{x}) \geq h(A\underline{\tilde{x}}) \tag{12}$$

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

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

1. $\underline{x}$ is Gaussian ($\underline{x} = \underline{\tilde{x}}$).

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 $\underline{x}$ 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 $\underline{x}$, and so (12) becomes

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

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

$$\frac{1}{m}h(A\underline{x}) \geq h(x). \tag{14}$$

As the dimension of $\underline{x}$ becomes large it may represent samples of a white stochastic process. In this case the matrix A represents linear transformation of that process. When A represents a filtering operation, some projections of $\underline{x}$ 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) results, in general, tighter bounds than the standard vector form EPI. Consider for example a vector $\underline{z} = A\underline{x} + B\underline{y}$ where both $\underline{x}, \underline{y}$ 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

$$P(\underline{z}) \geq P(A\underline{x}) + P(B\underline{y}) = |AP_xA^t|^{1/n} + |BP_yB^t|^{1/n} \tag{15}$$

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 $\underline{x}$ and $\underline{y}$ respectively. Our generalization of the EPI leads to the bound

$$P(\underline{z}) \geq P(A\underline{\tilde{x}} + B\underline{\tilde{y}}) = |AP_xA^t + BP_yB^t|^{1/n}. \tag{16}$$

By the Minkowski inequality (see e.g. Theorem 5 in [2]), the bound (16) is tighter than (15), with equality iff AP_xA^t is proportional to BP_yB^t.

2 Applications to Linear Transformations of a Vector with Independent Components

2.1 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 [3], 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. [4] pp. 231)

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

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 p.d.f'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

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

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 [5], pp. 20):

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

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 $\underline{x} = x_1 \ldots x_n$ with independent components. In stating this result we denote by $\underline{x}^*$ a Gaussian vector with independent components, such that $\mathrm{Var}\{x_i\}$ $=\mathrm{Var}\{x_i^*\}$. 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,*

$$\frac{1}{m}\mathcal{D}(A\underline{x};A\underline{x}^*) \leq \frac{1}{2}\log\left(\frac{|AR_xA^t|^{\frac{1}{m}}}{|AP_xA^t|^{\frac{1}{m}}}\right) \leq \max_{i=1\ldots n}\mathcal{D}(x_i;x_i^*) \tag{20}$$

where $m =$ Rank A, P_x is a diagonal matrix whose diagonal elements are $\{p_i\}$ the entropy powers of the components of $\underline{x}$ and R_x is the diagonal covariance matrix of $\underline{x}^*$ whose diagonal elements are $\{\sigma_i^2\}$, the powers of the components of $\underline{x}$.

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

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

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

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

Since $n \geq m = RankA$, the bound (21) is tighter.

2.2 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 $\underline{X} = [X_0, \ldots, X_{n-1}]^t$ be the DFT of $\underline{x} = [x_0, \ldots, x_{n-1}]^t$ i.e.

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

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 $I(X_0; X_1, \ldots, X_{n-1})$, i.e. the mutual information between the DC-component and the rest of the spectral components, has been considered in [6].

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

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

A similar definition can be made for $\mathcal{D}_h$. Now, in some applications r is fixed, while n becomes large, and so 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 [7], $\mathcal{D}_l \to 0$ as $n \to \infty$. A projection A_l for which $D_l \to 0$ as $n \to \infty$, 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

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

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 $\mathcal{D}(x_i; x_i^*) = 0.254$, 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 [2].

3 A Generalization of the Fisher-Information-Inequality

The duality between the EPI and various information inequalities has been pointed out in [2]. One example of such dual inequality is the Fisher-Information-Inequality (FII)

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

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

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

where ∇f is the n-dimensional gradient vector of f (see [2]). The scalar Fisher-information is defined as $J = \frac{1}{n}\mathrm{tr}\{K\} = \frac{1}{n}E\left\{\frac{1}{f^2}||\nabla f||^2\right\}$. The FII (25), whose proof is relatively simple, is actually used to prove the EPI (see [8] and [9]).

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

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

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

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

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 a TAU technical report, and here we sketch its structure. Similarly to the derivation in [8], [9] and [10], where the basic FII was shown, we show that

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

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

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

Using Cauchy-Schwarz inequality $EWW^t \geq EWEW^t$, it follows from (29) that

$$A^t\left(\frac{\nabla f(\underline{y})}{f(\underline{y})}\right)\left(\frac{\nabla f(\underline{y})}{f(\underline{y})}\right)^t A = E\left\{\frac{\nabla f(\underline{x})}{f(\underline{x})}\,\Big|\underline{y}\right\}E\left\{\frac{\nabla f(\underline{x})^t}{f(\underline{x})}\,\Big|\underline{y}\right\} \leq E\left\{\left(\frac{\nabla f(\underline{x})}{f(\underline{x})}\right)\left(\frac{\nabla f(\underline{x})}{f(\underline{x})}\right)^t\Big|\underline{y}\right\}. \tag{30}$$

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

$$A^tK(\underline{y})A \leq K(\underline{x}). \tag{31}$$

Finally, multiplying (31) from the left by $(AK(\underline{x})^{-1}A^t)^{-1}AK(\underline{x})^{-1}$, and 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 $A\underline{\hat{x}}$ 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 $K = J(x) \cdot I$, and if we further assume that A is orthonormal (i.e., $AA^t = I$), we can rewrite inequality (27) in a scalar form as $J(A\underline{x}) \leq J(x)$.

As in the standard EPI, one may hope that we can use the generalized FII to prove the generalized EPI, i.e. to prove Theorem 1. Indeed, as shown below, in the case where $\underline{x}$ is i.i.d., the generalized EPI can be proved via the generalized FII. Specifically, we use an integral relation between the divergence and the Fisher information given in [7] Lemma 1, following De-Bruijn's identity. In the vector case, this relation becomes

$$\mathcal{D}(\underline{y};\underline{y}^*) = \int_0^1 \mathrm{trace}\{R_y K(\underline{y}_t) - I\}\frac{dt}{2t} \tag{32}$$

where $\underline{y}_t = \sqrt{t}\,\underline{y} + \sqrt{1-t}\,\underline{y}^*$ ($\mathcal{D}$ here is measured in *nats*).

Applying (32) to $\underline{y} = A\underline{x}$ we get $R_y = AR_xA^t$. From the FII (27), $K(\underline{y}_t) = K(A\underline{x}_t) \leq (AK(\underline{x}_t)^{-1}A^t)^{-1}$. Now if $\underline{x}$ is i.i.d., then the components of $\underline{x}_t$ are also i.i.d., $R_x = \sigma^2 I$ and $K(\underline{x}_t) = J(x_t) \cdot I$. Incorporating into (32), we get

$$\mathcal{D}(A\underline{x};A\underline{x}^*) \leq \int_0^1 (\sigma^2 J(x_t) - 1) \cdot \mathrm{trace}\{I\}\frac{dt}{2t} = m\mathcal{D}(x;x^*) \tag{33}$$

where the second equality follows by applying (32) to the random variable x. Inequality (33) is equivalent to (13) and (21), i.e., to the generalization of the EPI in the i.i.d. case.

It seems plausible that the derivation above can be extended to the general case. We study this approach but at this point the proof of Theorem 1 via the double induction is still needed.

Acknowledgment

We thank Shlomo Shamai (Shitz) and the anonymous referee for pointing out the example in the end of Section 1.

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*. Figure 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 $\{a_{i,j}\}, i = 1 \ldots m,\ j = 1 \ldots n,$ the elements of the matrix A, and let 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 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

$$T = \begin{pmatrix} 1 & & & & \alpha_1 \\ & 1 & & 0 & \alpha_2 \\ & & \cdot & & \cdot \\ & & & \cdot & \cdot \\ & 0 & & \cdot & \cdot \\ & & & & 1 \end{pmatrix}, \qquad |T| = 1 \tag{34}$$

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

$$h(A\underline{x}) = h(T^{-1}\hat{A}\underline{x}) = h(\hat{A}\underline{x}) - \log|T| = h(\hat{A}\underline{x}). \tag{35}$$

Define some sub-matrices of $\hat{A}$, as follows

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

where $\underline{\hat{a}}_m = (a_{m,1}, \ldots, a_{m,n-1})^t$ is the last row of A without the last term, $\underline{\hat{b}}_n = (0, 0, \ldots, 0, a_{m,n})^t$ 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 $\underline{x}^- = x_1, \ldots, x_{n-1}$ 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, $\underline{\hat{a}}_m$, 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., $h(A^-\underline{x}^-) \geq h(A^-\underline{\tilde{x}}^-)$ and $h(B\underline{x}^-) \geq h(B\underline{\tilde{x}}^-)$.

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,

$$h(\hat{A}\underline{x}) = h(\hat{y}_1, \ldots, \hat{y}_m) = h(\hat{y}_1, \ldots, \hat{y}_{m-1}) + h(\hat{y}_m|\hat{y}_1, \ldots, \hat{y}_{m-1}) \tag{36}$$

and since $\hat{y}_m = \underline{a}_m^t\underline{x} = \underline{\hat{a}}_m^t\underline{x}^- + a_{m,n}x_n$ and $(\hat{y}_1, \ldots, \hat{y}_{m-1}) = A^-\underline{x}^-$, we can rewrite (36) as

$$h(\hat{A}\underline{x}) = h(A^-\underline{x}^-) + h(\underline{\hat{a}}_m^t\underline{x}^- + a_{m,n}x_n|A^-\underline{x}^-). \tag{37}$$

Notice that $a_{m,n}x_n$ in the RHS of (37) is independent of both $\underline{\hat{a}}_m^t\underline{x}^-$ and the condition $A^-\underline{x}^-$.

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

$$h(\hat{A}\underline{x}) = h(A^-\underline{x}^-) + h(a_{m,n}x_n) = h(A^-\underline{x}^-) + h(x_n) + \log|a_{m,n}| \quad . \tag{38}$$

Utilizing the induction assumption, asserting $h(A^-\underline{x}^-) \geq h(A^-\underline{\tilde{x}}^-)$, and by (35)

$$h(A\underline{x}) \geq h(A^-\underline{\tilde{x}}^-) + h(x_n) + \log|a_{m,n}| = h(A\underline{\tilde{x}}) \tag{39}$$

where the second equality follows by applying (38) to $h(A\underline{\tilde{x}})$ and since $h(x_n) = h(\tilde{x}_n)$. The induction step for this case is proved.

Consider now the second case where B has a full row-rank. Proceeding from (37), we use a conditional version of the EPI (originally presented in [9], see also [1] pp. 289) to lower bound the entropy of the sum of independent terms in the RHS of (37)

$$h(\hat{A}\underline{x}) \geq h(A^-\underline{x}^-) + \frac{1}{2}\log\left(2^{2h(\underline{\hat{a}}_m^t \underline{x}^- | A^-\underline{x}^-)} + 2^{2h(a_{m,n}x_n)}\right). \tag{40}$$

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

$$h(\hat{A}\underline{x}) \geq h(A^-\underline{x}^-) + \frac{1}{2}\log\left(2^{2[h(B\underline{x}^-) - h(A^-\underline{x}^-)]} + a_{m,n}^2 2^{2h(x_n)}\right). \tag{41}$$

The RHS of (41) 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 (41) is also monotonically increasing with $h(A^-\underline{x}^-)$. Since by the induction assumption $h(A^-\underline{x}^-) \geq h(A^-\underline{\tilde{x}}^-)$ and $h(B\underline{x}^-) \geq h(B\underline{\tilde{x}}^-)$, we can lower bound (41)

$$h(A\underline{x}) \geq h(A^-\underline{\tilde{x}}^-) + \frac{1}{2}\log\left(2^{2[h(B\underline{\tilde{x}}^-) - h(A^-\underline{\tilde{x}}^-)]} + a_{m,n}^2 2^{2h(x_n)}\right) \quad . \tag{42}$$

To complete the induction step, observe that the conditional version of the EPI used in the transition from (37) to (40) holds with equality for the Gaussian vector $\underline{\tilde{x}}$ and thus the RHS of (42) is $h(A\underline{\tilde{x}})$, as desired. □

Appendix B: Proof of Theorem 2

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

$$h(A\underline{x}) \geq h(A\underline{\tilde{x}}) = \frac{m}{2}\log(2\pi e|APA^t|^{\frac{1}{m}}) \tag{43}$$

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

$$h(A\underline{x}^*) - h(A\underline{x}) \leq h(A\underline{x}^*) - h(A\underline{\tilde{x}}) = \frac{m}{2}\log\left(\frac{|AR_xA^t|^{\frac{1}{m}}}{|APA^t|^{\frac{1}{m}}}\right). \tag{44}$$

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

$$\frac{1}{m}\mathcal{D}(A\underline{x}; A\underline{x}^*) \leq \frac{1}{2}\log\left(\frac{|AR_xA^t|^{\frac{1}{m}}}{|APA^t|^{\frac{1}{m}}}\right) \quad . \tag{45}$$

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

$$\frac{1}{m}\mathcal{D}(A\underline{x}; A\underline{x}^*) \leq \frac{1}{2}\log\left(\frac{\sigma^2}{p}\right) = \mathcal{D}(x; x^*), \tag{46}$$

which proves (21).

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

$$\frac{|AR_xA^t|^{\frac{1}{m}}}{|APA^t|^{\frac{1}{m}}} \leq \max_{i=1\ldots n}\frac{\sigma_i^2}{p_i} \tag{47}$$

and the second inequality in (20) follows since $\mathcal{D}(x_i; x_i^*) = \frac{1}{2}\log\left(\frac{\sigma_i^2}{p_i}\right)$. 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., Rank $A = m$ is less then the number of rows. Using the more general definition of the divergence given in (19), it follows that if $\underline{y}_1 = T\underline{y}$, i.e., if $\underline{y}_1$ linearly depends on $\underline{y}$, then

$$\mathcal{D}(\underline{y}, \underline{y}_1; \underline{y}^*, \underline{y}_1^*) = \mathcal{D}(\underline{y}, T\underline{y}; \underline{y}^*, T\underline{y}^*) = \mathcal{D}(\underline{y}, \underline{y}^*). \tag{48}$$

Now, the vector $(A\underline{x})$ can be separated into $(A_o\underline{x}, A_+\underline{x})$ where the $m \times n$ matrix A_o has a full row-rank and the augmented part, $A_+\underline{x}$, linearly depends on $A_o\underline{x}$. Thus, by (48)

$$\mathcal{D}(A\underline{x}; A\underline{x}^*) = \mathcal{D}(A_o\underline{x}; A_o\underline{x}^*), \tag{49}$$

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

$$\frac{|A\Lambda A^t|^{\frac{1}{m}}}{|APA^t|^{\frac{1}{m}}} \leq \max_{i=1\ldots n} \frac{\lambda_i}{p_i} \tag{50}$$

Proof: Define

$$r_m \triangleq \max_{i=1\ldots n} \frac{\lambda_i}{p_i}\ . \tag{51}$$

Clearly, $r_m \cdot p_i - \lambda_i \geq 0$ for any $i = 1 \ldots n$, 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

$$A(r_m \cdot P - \Lambda)A^t \geq 0 \quad \Longrightarrow \quad 0 \leq A\Lambda A^t \leq r_m \cdot APA^t\ \ . \tag{52}$$

The inequality (52) 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)

$$|A\Lambda A^t| \leq |r_m APA^t| = (r_m)^m |APA^t| \tag{53}$$

and (50) is proved. $\square$

Appendix C: Proof of Theorem 3

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

$$I(\underline{y}_l; \underline{y}_h) = I(\underline{y}_l^*; \underline{y}_h^*) - \mathcal{D}(\underline{y}_l; \underline{y}_l^*) - \mathcal{D}(\underline{y}_h; \underline{y}_h^*) + \mathcal{D}(\underline{y}_l, \underline{y}_h; \underline{y}_l^*, \underline{y}_h^*). \tag{54}$$

Examine now each term in the RHS of (39). Since orthogonality implies independence for zero-mean Gaussian vectors,

$$I(\underline{y}_l^*; \underline{y}_h^*) = 0 \ . \tag{55}$$

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

$$\mathcal{D}(\underline{y}_h; \underline{y}_h^*) \leq (n-r)\mathcal{D}(x; x^*). \tag{56}$$

Finally,

$$\mathcal{D}(\underline{y}_l, \underline{y}_h; \underline{y}_l^*, \underline{y}_h^*) = \mathcal{D}(\underline{y}; \underline{y}^*) = \mathcal{D}(\underline{x}; \underline{x}^*) = n\mathcal{D}(x; x^*) \tag{57}$$

since A_l and A_h compose together an invertible transformation which preserves the divergence. Combining (23) and (54)-(57) yields the desired result. $\Box$

References

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

[2] A. Dembo, T.M.Cover, and J.A.Thomas. Information theoretic inequalities. *IEEE Trans. Information Theory*, IT-37:1501–1518, Nov. 1991.

[3] D. Donoho. On minimum entropy deconvolution. *Applied Time Series Analysis II*, pages 565–608, Academic Press, NY, 1981.

[4] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* Wiley, New York, 1991.

[5] M. S. Pinsker. *Information and Information Stability of Random Variables and Processes.* Holden Day, San Francis. CA., 1964.

[6] R. Zamir and M. Feder. Rate distortion performance in coding band-limited sources by sampling and dithered quantization. *IEEE Trans. Information Theory*, IT-41:141–154, Jan. 1995.

[7] A.R. Barron. Entropy and the central limit theorem. *The Annals of Probability*, 14, No. 1:336–342, 1986.

[8] A. J. Stam. Some inequalities satisfied by the quantities of information of Fisher and Shannon. *Inform. Control*, 2:101–112, June 1959.

[9] N. M. Blachman. The convolution inequality for entropy powers. *IEEE Trans. Information Theory*, IT-11:267–271, 1965.

[10] A. Dembo. Simple Proof of the Concavity of the Entropy Power with respect to Added Gaussian Noise. *IEEE Trans. Information Theory*, IT-35:887–888, July 1989.