HAL
open science

A simple proof of the entropy-power inequality via properties of mutual information

Olivier Rioul

To cite this version:

Olivier Rioul. A simple proof of the entropy-power inequality via properties of mutual information. 2007 IEEE International Symposium on Information Theory (ISIT 2007), Jun 2007, Nice, France. 10.1109/ISIT.2007.4557202 . hal-03329762

HAL Id: hal-03329762

https://telecom-paris.hal.science/hal-03329762v1

Submitted on 10 Aug 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A Simple Proof of the Entropy-Power Inequality via Properties of Mutual Information

Olivier Rioul
Dept. ComElec, GET/Télécom Paris (ENST)
ParisTech Institute & CNRS LTCI
Paris, France
Email: olivier.rioul@enst.fr

***Abstract*— While most useful information theoretic inequalities can be deduced from the basic properties of entropy or mutual information, Shannon's entropy power inequality (EPI) seems to be an exception: available information theoretic proofs of the EPI hinge on integral representations of differential entropy using either Fisher's information (FI) or minimum mean-square error (MMSE). In this paper, we first present a unified view of proofs via FI and MMSE, showing that they are essentially dual versions of the same proof, and then fill the gap by providing a new, simple proof of the EPI, which is solely based on the properties of mutual information and sidesteps both FI or MMSE representations.**

I. Introduction

Shannon's entropy power inequality (EPI) gives a lower bound on the differential entropy of the sum of independent random variables X, Y with densities:

$$\exp(2h(X+Y)) \geq \exp(2h(X)) + \exp(2h(Y)) \qquad (1)$$

with equality if X and Y are Gaussian random variables. The differential entropy of the probability density function $p(x)$ of X is defined as

$$h(X) = \mathsf{E}\Big\{\log \frac{1}{p(X)}\Big\}, \qquad (2)$$

where it is assumed throught this paper that all logarithms are natural.

The EPI finds its application in proving converses of channel or source coding theorems. It was used by Shannon as early as his 1948 paper [1] to bound the capacity of non-Gaussian additive noise channels. Recently, it was used to determine the capacity region of the Gaussian MIMO broadcast channel [2]. The EPI also finds application in blind source separation and deconvolution (see, e.g., [3]) and is instrumental in proving a strong version of the central limit theorem with convergence in relative entropy [4].

Shannon's proof of the EPI [1] was incomplete in that he only checked that the necessary condition for a local minimum of $h(X+Y)$ is satisfied. Available rigorous proofs of the EPI are in fact proofs of an alternative statement

$$h(\sqrt{\lambda}\,X + \sqrt{1-\lambda}\,Y) \geq \lambda h(X) + (1-\lambda)h(Y) \qquad (3)$$

for any $0 \leq \lambda \leq 1$, which amounts to the concavity of the entropy under the "variance preserving" transformation [5]:

$$(X,Y) \longmapsto W = \sqrt{\lambda}\,X + \sqrt{1-\lambda}\,Y. \qquad (4)$$

To see that (3) is equivalent to (1), define U, V by the relations $X = \sqrt{\lambda}\,U$, $Y = \sqrt{1-\lambda}\,V$, and rewrite (1) as follows:

$$e^{2h(\sqrt{\lambda}\,U + \sqrt{1-\lambda}\,V)} \geq \lambda e^{2h(U)} + (1-\lambda)e^{2h(V)}.$$

Taking logarithms of both sides, (3) follows from the concavity of the logarithm. Conversely, taking exponentials, (3) written for U, V implies (1) for λ chosen so that U and V have equal entropies: $\exp 2h(U) = \exp 2h(V)$, that is, $\frac{\exp 2h(X)}{\lambda} = \frac{\exp 2h(Y)}{1-\lambda}$ or $\lambda = e^{2h(X))}/(e^{2h(X)} + e^{2h(Y)})$.

The first rigorous proof of the EPI was given by Stam [6] (see also Blachman [7]). It is based on the properties of Fisher's information (FI)

$$J(X) = \mathsf{E}\Big\{\Big(\frac{p'(X)}{p(X)}\Big)^2\Big\}, \qquad (5)$$

the link between differential entropy and FI being de Bruijn's identity [8, Thm. 17.7.2]:

$$\frac{d}{dt}h(X+\sqrt{t}\,Z) = \frac{1}{2}J(X+\sqrt{t}\,Z), \qquad (6)$$

where $Z \sim \mathcal{N}(0,1)$ is a standard Gaussian random variable, which is independent of X. Recently, Verdú, Guo and Shamai [9], [10] provided an alternative proof of the EPI based on the properties of the MMSE in estimating the input X to a Gaussian channel given the output $Y = \sqrt{t}\,X + Z$, where t denotes the signal-to-noise ratio. This MMSE is achieved by the conditional mean estimator $\hat{X}(Y) = \mathsf{E}(X|Y)$ and is given by the conditional variance

$$\mathsf{Var}(X|Y) = \mathsf{E}\Big\{(X - \mathsf{E}\{X|Y\})^2\Big\}, \qquad (7)$$

where the expectation is taken over the joint distribution of the random variables X and Y. The connection between input-output mutual information $I(X;Y) = h(Y) - h(Z)$ and MMSE is made by the following identity derived in [11]:

$$\frac{d}{dt}I(X;\sqrt{t}\,X+Z) = \frac{1}{2}\mathsf{Var}(X|\sqrt{t}\,X+Z). \qquad (8)$$

This identity turns out to be equivalent to de Bruijn's identity (6). It has been claimed [10] that using the alternative MMSE representation in place of FI representation is more insightful and convenient for proving the EPI.

In this paper, we show that it is possible to avoid both MMSE and FI representations and use only basic properties

of mutual information. The new proof of the EPI presented in this paper is based on a convexity inequality for mutual information under the variance preserving transformation (4):

Theorem 1: If X and Y are independent random variables, and if Z is Gaussian independent of X, Y, then

$$I(\sqrt{\lambda}\,X + \sqrt{1-\lambda}\,Y + \sqrt{t}\,Z; Z) \leq \lambda I(X + \sqrt{t}\,Z; Z) + (1-\lambda) I(Y + \sqrt{t}\,Z; Z) \quad (9)$$

for all $0 \leq \lambda \leq 1$ and $t \geq 0$.

Apart from its intrinsic interest, we show that inequality (9) reduces to the EPI by letting $t \to \infty$.

Before turning to the proof of Theorem 1, we make the connection between earlier proofs of the EPI via FI and via MMSE by focusing on the essential ingredients common to the proofs. This will give an idea of the level of difficulty that is required to understand the conventional approaches, while also serving as a guide to understand the new proof which uses similar ingredients, but is comparatively simpler and shorter.

The remainder of the paper is organized as follows. Section II gives a direct proof of a simple relation between FI and MMSE, interprets (6) and (8) as dual consequences of a generalized identity, and explores the relationship between the two previous proofs of the EPI via FI and via MMSE. It is shown that these are essentially dual versions of the same proof; they follow the same lines of thought and each step has an equivalent formulation, and a similar interpretation, in terms of FI and MMSE. Section III then proves Theorem 1 and the EPI using two basic ingredients common to earlier approaches, namely 1) a "data processing" argument applied to (4); 2) a Gaussian perturbation method. The reader may wish to skip to this section first, which does not use the results presented earlier. The new approach has the advantage of being very simple in that it relies only on the basic properties of mutual information.

II. Proofs of the EPI via FI and MMSE revisited

The central link between FI and MMSE takes the form of a simple relation which shows that they are complementary quantities in the case of a standard Gaussian perturbation Z independent of X:

$$J(X+Z) + \mathsf{Var}(X|X+Z) = 1 \quad (10)$$

This identity was mentioned in [11] to show that (6) and (8) are equivalent. We first provide a direct proof of this relation, and then use it to unify and simplify existing proofs of the EPI via FI and via MMSE. In particular, two essential ingredients, namely, Fisher's information inequality [7], [12], and a related inequality for MMSE [9], [10], will be shown to be equivalent from (10).

A. A new proof of (10)

Fisher's information (5) can be written in the form

$$J(X) = \mathsf{E}\{S^2(X)\} = \mathsf{Var}\{S(X)\} \quad (11)$$

where $S(X) = p'(X)/p(X)$ is a zero-mean random variable. The following conditional mean representation is due to Blachman [7]:

$$S(X+Z) = \mathsf{E}\{S(Z)|X+Z\}. \quad (12)$$

By the "law of total variance", this gives

$$\begin{aligned} J(X+Z) &= \mathsf{Var}\{S(X+Z)\} \\ &= \mathsf{Var}\{\mathsf{E}\{S(Z)|X+Z\}\} \\ &= \mathsf{Var}\{S(Z)\} - \mathsf{Var}\{S(Z)|X+Z\} \\ &= J(Z) - \mathsf{Var}\{S(Z)|X+Z\} \end{aligned} \quad (13)$$

We now use the fact that Z is standard Gaussian. It is easily seen by direct calculation that $S(Z) = -Z$ and $J(Z) = 1$, and, therefore, $J(X+Z) = 1 - \mathsf{Var}\{Z|X+Z\}$. Since $Z - \mathsf{E}(Z|X+Z) = \mathsf{E}(X|X+Z) - X$ we have $\mathsf{Var}\{Z|X+Z\} = \mathsf{Var}\{X|X+Z\}$, thereby showing (10). ■

Note that when Z is Gaussian but not standard Gaussian, we have $S(Z) = -Z/\mathsf{Var}(Z)$ and $J(Z) = 1/\mathsf{Var}(Z)$, and (10) generalizes to

$$\mathsf{Var}(Z)J(X+Z) + J(Z)\mathsf{Var}(X|X+Z) = 1. \quad (14)$$

Another proof, which is based on a data processing argument and avoids Blachman's representation, is given in [13].

B. Intepretation

Equation (10) provides a new estimation theoretic interpretation of Fisher's information of a noisy version $X' = X + Z$ of X. It is just the complementary quantity to the MMSE that results from estimating X from X'. The estimation is all the more better as the MMSE is lower, that is, as X' provides higher FI. Thus Fisher's information is a measure of least squares estimation's efficiency, when estimation is made in additive Gaussian noise.

To illustrate, consider the special case of a Gaussian random variable X. Then the best estimator is the linear regression estimator, with MMSE equal to $\mathsf{Var}(X|X') = (1-\rho^2)\mathsf{Var}(X)$ where $\rho = \sqrt{\mathsf{Var}(X)/\mathsf{Var}(X')}$ is the correlation factor between X and X':

$$\mathsf{Var}(X|X') = \frac{\mathsf{Var}(X)}{\mathsf{Var}(X)+1}. \quad (15)$$

Meanwhile, $J(X')$ is simply the reciprocal of the variance of X':

$$J(X') = \frac{1}{\mathsf{Var}(X)+1}. \quad (16)$$

Both quantities sum to one, in accordance with (10). In the case of non-Gaussian noise, we have the more general identity (13) which also links Fisher information and conditional variance, albeit in a more complicated form.

C. Dual versions of de Bruijn's Identity

De Bruijn's identity can be stated in the form [5]

$$\frac{d}{dt} h(X + \sqrt{t}\,Z)\Big|_{t=0} = \frac{1}{2} J(X)\mathsf{Var}(Z). \quad (17)$$

The conventional technical proof of (17) is obtained by integrating by parts using a diffusion equation satisfied by the

Gaussian distribution (see e.g., [7] and [8, Thm. 17.7.2]). A simpler proof of a more general result is included in [13]. From (17), we deduce the following.

Theorem 2 (de Bruijn's identity): For any two random independent random variables X and Z,

$$\frac{d}{dt}h(X+\sqrt{t}\,Z) = \frac{1}{2}\,J(X+\sqrt{t}\,Z)\,\mathsf{Var}(Z) \tag{18}$$

if Z is Gaussian, and

$$\frac{d}{dt}h(X+\sqrt{t}\,Z) = \frac{1}{2}\,J(X)\,\mathsf{Var}(Z|X+\sqrt{t}\,Z) \tag{19}$$

if X is Gaussian.

This theorem is essentially contained in [11]. In fact, noting that $I(X;\sqrt{t}\,X+Z) = h(\sqrt{t}\,X+Z) - h(Z)$, it is easily seen that (19), with X and Z interchanged, is the identity (8) of Guo, Verdú and Shamai. Written in the form (19) it is clear that this is a dual version of the conventional de Bruijn's identity (18). Note that both identities reduce to (17) for $t=0$. For completeness we include a simple proof for $t>0$.

Proof: Equation (18) easily follows from (17) using the stability property of Gaussian distributions under convolution: substitute $X+\sqrt{t'}\,Z'$ for X in (17), where Z and Z' are taken to be iid Gaussian random variables, and use the fact that $\sqrt{t}\,Z+\sqrt{t'}\,Z'$ and $\sqrt{t+t'}\,Z$ are identically distributed.

To prove (19) we use the complementary relation (14) in the form

$$\mathsf{Var}(X)J(X+\sqrt{t}\,Z) + tJ(X)\mathsf{Var}(Z|X+\sqrt{t}\,Z) = 1 \tag{20}$$

where X is Gaussian. Let $u=1/t$. By (18) (with X and Z interchanged), we have

$$\begin{aligned}\frac{d}{dt}h(X+\sqrt{t}\,Z) &= \frac{d}{dt}\Big\{h(\sqrt{u}\,X+Z)+\frac{1}{2}\log t\Big\}\\ &= -\frac{1}{t^2}\frac{d}{du}h(\sqrt{u}\,X+Z)+\frac{1}{2t}\\ &= -\frac{1}{2t^2}\mathsf{Var}(X)J(\sqrt{u}\,X+Z)+\frac{1}{2t}\\ &= -\frac{1}{2t}\mathsf{Var}(X)J(X+\sqrt{t}\,Z)+\frac{1}{2t}.\end{aligned}$$

which combined with (20) proves (19). ■

D. Equivalent integral representations of differential entropy

Consider any random variable with density and finite variance $\sigma^2 = \mathsf{Var}(X)$. Its non-Gaussianness is defined as the divergence with respect to a Gaussian random variable X_G with identical second centered moments, and is given by

$$D_h(X) = h(X_G) - h(X) \tag{21}$$

where $h(X_G) = \frac{1}{2}\log\big(2\pi e\sigma^2\big)$. Let Z be standard Gaussian, independent of X. From (18), we obtain

$$\frac{d}{dt}D_h(X+\sqrt{t}\,Z) = -\frac{1}{2}D_J(X+\sqrt{t}\,Z), \tag{22}$$

where

$$D_J(X) = J(X) - J(X_G). \tag{23}$$

Here $J(X_G) = 1/\sigma^2$ and (23) is nonnegative by the Cramér-Rao inequality. Now for $t=0$, $D_h(X+\sqrt{t}\,Z) = D_h(X)$, and since non-Gaussianness is scale invariant, $D_h(X+\sqrt{t}\,Z) = D_h(Z+X/\sqrt{t}) \to D_h(Z) = 0$ as $t\to+\infty$. Therefore, integrating (22) from $t=0$ to $+\infty$ we obtain a FI integral representation for differential entropy [4]

$$D_h(X) = \frac{1}{2}\int_0^\infty D_J(X+\sqrt{t}\,Z)\,dt \tag{24}$$

or

$$h(X) = \frac{1}{2}\log\big(2\pi e\sigma^2\big) - \frac{1}{2}\int_0^\infty J(X+\sqrt{t}\,Z) - \frac{1}{\sigma^2+t}\,dt. \tag{25}$$

Similarly, from (19) with X and Z interchanged, we obtain a dual identity:

$$\frac{d}{dt}D_h(\sqrt{t}\,X+Z) = \frac{1}{2}D_V(X|\sqrt{t}\,X+Z), \tag{26}$$

where for $Y=\sqrt{t}\,X+Z$ and $Y_G=\sqrt{t}\,X_G+Z$,

$$D_V(X|Y) = \mathsf{Var}(X_G|Y_G) - \mathsf{Var}(X|Y). \tag{27}$$

Again this quantity is nonnegative, because $\mathsf{Var}(X_G|Y_G) = \sigma^2/(t\sigma^2+1)$ is the MMSE achievable by a linear estimator, which is suboptimal for non-Gaussian X. Now for $t=0$, $D_h(\sqrt{t}\,X+Z) = D_h(Z)$ vanishes, and since non-Gaussianness is scale invariant, $D_h(\sqrt{t}\,X+Z) = D_h(X+Z/\sqrt{t}) \to D_h(X)$ as $t\to+\infty$. Therefore, integrating (26) from $t=0$ to $+\infty$ we readily obtain the MMSE integral representation for differential entropy [11]:

$$D_h(X) = \frac{1}{2}\int_0^\infty D_V(X|\sqrt{t}\,X+Z)\,dt \tag{28}$$

or

$$h(X) = \frac{1}{2}\log\big(2\pi e\sigma^2\big) - \frac{1}{2}\int_0^\infty \frac{\sigma^2}{1+t\sigma^2} - \mathsf{Var}(X|\sqrt{t}\,X+Z)\,dt. \tag{29}$$

This MMSE representation was first derived in [11] and used in [9], [10] to prove the EPI. The FI representation (25) can be used similarly, yielding essentially Stam's proof of the EPI [6], [7]. These proofs are sketched below.

Note that the equivalence between FI and MMSE representations (24), (28) is immediate by the complementary relation (10), which can be simply written as

$$D_J(X+Z) = D_V(X|X+Z). \tag{30}$$

In fact, it is easy to see that (24) and (28) prove each other by (30) after making a change of variable $u=1/t$. Both sides of (30) are of course simultaneously nonnegative and measure a "non-Gaussianness" of X when estimated in additive Gaussian noise. Interestingly, these FI and MMSE non-Gaussianities coincide.

It is also useful to note that in the above derivations, the Gaussian random variable X_G may very well be chosen such that $\sigma^2 = \mathsf{Var}(X_G)$ is not equal to $\mathsf{Var}(X)$. Formulas (21)–(30) still hold, even though quantities such as D_h, D_J, and D_V may take negative values. In particular, the right-hand sides of (25) and (29) do not depend on the particular choice of σ.

E. Simplified proofs of the EPI via FI and via MMSE

For Gaussian random variables X_G, Y_G with the same variance, the EPI in the form (3) holds trivially with equality. Therefore, to prove the EPI, it is sufficient to show the following convexity property:

$$D_h(W) \leq \lambda D_h(X) + (1-\lambda) D_h(Y) \tag{31}$$

where $D_h(\cdot)$ is defined by (21) and W is defined as in (4). By the last remark of the preceding subsection, the FI and MMSE representations (24), (28) hold. Therefore, to prove the EPI, it is sufficient to show that either one of the following inequalities holds.

$$\begin{aligned} &D_J(W+\sqrt{t}\,Z) \\ &\qquad \leq \lambda D_J(X+\sqrt{t}\,Z) + (1-\lambda) D_J(Y+\sqrt{t}\,Z) \\ &D_V(W|\sqrt{t}\,W+Z) \\ &\qquad \leq \lambda D_V(X|\sqrt{t}\,X+Z) + (1-\lambda) D_V(Y|\sqrt{t}\,Y+Z). \end{aligned}$$

These in turn can be written as

$$\begin{aligned} &J(W+\sqrt{t}\,Z) \\ &\qquad \leq \lambda J(X+\sqrt{t}\,Z) + (1-\lambda) J(Y+\sqrt{t}\,Z) \qquad (32) \\ &\mathsf{Var}(W|\sqrt{t}\,W+Z) \\ &\qquad \geq \lambda \mathsf{Var}(X|\sqrt{t}\,X+Z) + (1-\lambda) \mathsf{Var}(Y|\sqrt{t}\,Y+Z), \qquad (33) \end{aligned}$$

because these inequalities hold trivially with equality for X_G and Y_G.

Inequality (33) is easily proved in the form

$$\mathsf{Var}(W|\sqrt{t}\,W+Z) \geq \mathsf{Var}(W|\sqrt{t}\,X+Z', \sqrt{t}\,Y+Z'') \tag{34}$$

where Z' and Z'' are standard Gaussian random variables, independent of X, Y and of each other, and $Z = \sqrt{\lambda}\,Z' + \sqrt{1-\lambda}\,Z''$. This has a simple interpretation [9], [10]: it is better to estimate the sum of independent random variables from individual noisy measurements than from the sum of these measurements.

The dual inequality (32) can also be proved in the form

$$J(W) \leq \lambda J(X) + (1-\lambda) J(Y) \tag{35}$$

where we have substituted X, Y for $X+\sqrt{t}\,Z', Y+\sqrt{t}\,Z''$. This inequality is known as the Fisher's information inequality [5], and an equivalent formulation is

$$\frac{1}{J(X+Y)} \geq \frac{1}{J(X)} + \frac{1}{J(Y)}. \tag{36}$$

Blachman [7] gave a short proof using representation (12) and Zamir [12] gave an insightful proof using a data processing argument. Again (36) has a simple interpretation, which is very similar to that of (34). Here $1/J(X)$ is the Cramér-Rao lower bound (CRB) of the mean-squared error of the unbiaised estimator X for a translation parameter, and (36) states that in terms of the CRB it is better to estimate the translation parameter corresponding to the sum of independent random variables $X+Y$ from individual measurements than from the sum of these measurements.

At any rate, the equivalence between (32) and (33) is immediate by the complementary relation (10) or its generalized version (14), as can be easily seen. Either one of (32), (33) gives a proof of the EPI. ■

The above derivations illuminate intimate connections between both proofs of the EPI, via FI and via MMSE. They do not only follow the same lines of argumentation, but are also shown to be dual in the sense that through (10), each step in these proofs has an equivalent formulation and similar interpretation in terms of FI or MMSE.

III. A NEW PROOF OF THE EPI

For convenience in the following proof, define W by (4) and let $\alpha = \sqrt{t\lambda}$ and $\beta = \sqrt{t(1-\lambda)}$.

A. Proof of Theorem 1

Similarly as in the conventional proofs of the EPI, we use the fact that the linear combination $\sqrt{\lambda}(X+\alpha Z) + \sqrt{1-\lambda}(Y+\beta Z) = W + \sqrt{t}\,Z$ cannot bring more information than the individual variables $X+\alpha Z$ and $Y+\beta Z$ together. Thus, by the data processing inequality for mutual information,

$$I(W+\sqrt{t}\,Z; Z) \leq I(X+\alpha Z, Y+\beta Z; Z). \tag{37}$$

Let $U = X+\alpha Z$, $V = Y+\beta Z$ and develop using the chain rule for mutual information:

$$\begin{aligned} I(U,V;Z) &= I(U;Z) + I(V;Z|U) \\ &\leq I(U;Z) + I(V;Z|U) + I(U;V) \\ &= I(U;Z) + I(V;U,Z) \\ &= I(U;Z) + I(V;Z) + I(U;V|Z). \end{aligned}$$

Since X and Y are independent, U and V are conditionally independent given Z, and therefore, $I(U;V|Z) = 0$. Thus, we obtain the inequality

$$I(W+\sqrt{t}\,Z; Z) \leq I(X+\alpha Z; Z) + I(Y+\beta Z; Z) \tag{38}$$

Assume for the moment that $I(X+\alpha Z; Z)$ admits a second-order Taylor expansion about $\alpha = 0$ as $t \to 0$. Since $I(X+\alpha Z; Z)$ vanishes for $\alpha = 0$, and since mutual information is nonnegative, we may write

$$I(X+\alpha Z; Z) = \lambda I(X+\sqrt{t}\,Z; Z) + o(\alpha^2)$$

where $o(\alpha^2) = o(t)$ and $\frac{o(t)}{t}$ tends to zero as $t \to 0$. Similarly $I(Y+\beta Z; Z) = (1-\lambda) I(Y+\sqrt{t}\,Z; Z) + o(t)$. It follows that in the vicinity of $t = 0$,

$$\begin{aligned} I(W+\sqrt{t}\,Z; Z) &\leq \lambda I(X+\sqrt{t}\,Z; Z) \\ &\quad + (1-\lambda) I(Y+\sqrt{t}\,Z; Z) + o(t). \end{aligned} \tag{39}$$

We now remove the $o(t)$ term by using the assumption that Z is Gaussian. Consider the variables $X' = X + \sqrt{t'}\,Z_1'$, $Y' = Y + \sqrt{t'}\,Z_2'$, where Z_1', Z_2' are identically distributed as Z but independent of all other random variables. This Gaussian perturbation ensures that densities are smooth, so that $I(X'+\alpha Z; Z)$ and $I(Y'+\beta Z; Z)$ both admit a second-order Taylor

expansion about $t = 0$. We may, therefore, apply (39) to X' and Y', which gives

$$I(W + \sqrt{t'}\, Z' + \sqrt{t}\, Z; Z) \leq \lambda I(X + \sqrt{t'}\, Z_1' + \sqrt{t}\, Z; Z) + (1-\lambda) I(Y + \sqrt{t'}\, Z_2' + \sqrt{t}\, Z; Z) + o(t) \quad (40)$$

where $Z' = \sqrt{\lambda}\, Z_1' + \sqrt{1-\lambda}\, Z_2'$ is identically distributed as Z. Applying the obvious identity $I(X+Z'+Z; Z) = I(X+Z'+Z; Z'+Z) - I(X+Z'; Z')$ and using the stability property of the Gaussian distribution under convolution, (40) boils down to

$$f(t'+t) \leq f(t') + o(t)$$

where we have noted $f(t) = I(W + \sqrt{t}\, Z; Z) - \lambda I(X + \sqrt{t}\, Z; Z) - (1-\lambda) I(Y + \sqrt{t}\, Z; Z)$. It follows that $f(t)$ is non increasing in t, and since it clearly vanishes for $t = 0$, it always assumes non positive values for all $t \geq 0$. This completes the proof of (9). ■

Interestingly, this proof uses two basic ingredients common to earlier proofs presented in section II: 1) the fact that two variables together bring more information than their sum, which is here expressed as a data processing inequality for mutual information; 2) a Gaussian perturbation argument using an auxiliary variable Z.

In fact, Theorem 1 could also be proved using either one of the integral representations (25), (29), which are equivalent by virtue of (10) and obtained through de Bruijn's identity as explained in section II. The originality in the present proof is that it does neither require de Bruijn's identity nor the notions of FI or MMSE.

B. The discrete case

The above proof of Theorem 1 does not require that X or Y be random variables with densities. Therefore, Theorem 1 also holds for discrete (finitely or countably) real valued random variables. Verdú and Guo [9] proved that the EPI, in the form (3), also holds in this case, where differential entropies are replaced by entropies. We call attention that this is in fact a trivial consequence of the stronger inequality

$$H(\sqrt{\lambda}\, X + \sqrt{1-\lambda}\, Y) \geq \max\big(H(X), H(Y)\big) \quad (41)$$

for any two independent discrete random variables X and Y. This inequality is easily obtained by noting that

$$H(W) \geq H(W|Y) = H(X|Y) = H(X) \quad (42)$$

and similarly for $H(Y)$.

C. Proof of the EPI

We now show that the EPI for differential entropies, in the form (3), follows from Theorem 1. By the identity $I(X + Z; Z) + h(X) = I(X; X + Z) + h(Z)$, inequality (9) can be written in the form

$$h(W) - \lambda h(X) - (1-\lambda) h(Y) \geq I(W; W + \sqrt{t}\, Z) - \lambda I(X; X + \sqrt{t}\, Z) - (1-\lambda) I(Y; Y + \sqrt{t}\, Z). \quad (43)$$

We now let $t \to \infty$ in the right-hand side of this inequality. Let $\varepsilon = 1/\sqrt{t}$ and X_G be a Gaussian random variable independent of Z, with identical second moments as X. Then $I(X; X + \sqrt{t}\, Z) = I(X; \varepsilon X + Z) = h(\varepsilon X + Z) - h(Z) \leq h(\varepsilon X_G + Z) - h(Z) = \frac{1}{2}\log(1 + \sigma_X^2/t\sigma_Z^2)$, which tends to zero as $t \to \infty$. This holds similarly for the other terms in the right-hand side of (43). Therefore, the EPI (3) follows. ■

In light of this proof, we see that theorem 1, which contains the EPI as the special case where $\sigma_Z^2 \to \infty$, merely states that the difference $h(W + Z) - \lambda h(X + Z) - (1-\lambda) h(Y + Z)$ between both sides of the EPI (3) decreases as independent Gaussian noise Z is added. This holds in accordance with the fact that this difference is zero for Gaussian random variables with identical variances.

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 and the EPI, in a similar manner as Verdú and Guo's proof uses (29) rather than (25). In fact, this would amount to proving (43), whose natural derivation using the data processing inequality for mutual information is through (37).

The same proof we used above can be employed verbatim to prove the EPI for random vectors. Generalizations to various extended versions of EPI are provided in a follow-up to this work [13].

References

[1] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 623–656, Oct. 1948.

[2] H. Weingarten, Y. Steinberg, and S. Shamai (Shitz), "The capacity region of the Gaussian multiple-input multiple-output broadcast channel," *IEEE Transactions on Information Theory*, vol. 52, no. 9, pp. 3936–3964, Sept. 2006.

[3] J. F. Bercher and C. Vignat, "Estimating the entropy of a signal with applications," *IEEE Transactions on Signal Processing*, vol. 48, no. 6, pp. 1687–1694, June 2000.

[4] A. R. Barron, "Entropy and the central limit theorem," *The Annals of Probability*, vol. 14, no. 1, 336–342 1986.

[5] A. Dembo, T. M. Cover, and J. A. Thomas, "Information theoretic inequalities," *IEEE Transactions on Information Theory*, vol. 37, no. 6, pp. 1501–1518, Nov. 1991.

[6] A. J. Stam, "Some inequalities satisfied by the quantities of information of Fisher and Shannon," *Information and Control*, vol. 2, pp. 101–112, June 1959.

[7] N. M. Blachman, "The convolution inequality for entropy powers," *IEEE Transactions on Information Theory*, vol. 11, no. 2, pp. 267–271, Apr. 1965.

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

[9] S. Verdú and D. Guo, "A simple proof of the entropy-power inequality," *IEEE Transactions on Information Theory*, vol. 52, no. 5, pp. 2165–2166, May 2006.

[10] D. Guo, S. Shamai (Shitz), and S. Verdú, "Proof of entropy power inequalities via MMSE," in *Proceedings of the IEEE International Symposium on Information Theory*, Seattle, USA, July 2006, pp. 1011–1015.

[11] ——, "Mutual information and minimum mean-square error in Gaussian channels," *IEEE Transactions on Information Theory*, vol. 51, no. 4, pp. 1261–1282, Apr. 2005.

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

[13] O. Rioul, "Information theoretic proofs of entropy power inequalities," *IEEE Transactions on Information Theory*, submitted for publication, arXiv cs.IT/0704.1751, April 2007.