

Optimal transportation to the entropy-power inequality

Olivier Rioul

▶ **To cite this version:**

Olivier Rioul. Optimal transportation to the entropy-power inequality. IEEE Information Theory and Applications Workshop (ITA 2017), Feb 2017, San Diego, United States. 10.1109/ITA.2017.8023467 . hal-02288488

HAL Id: hal-02288488

https://telecom-paris.hal.science/hal-02288488

Submitted on 12 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.

Optimal Transportation to the Entropy-Power Inequality

Olivier Rioul
LTCI, Télécom ParisTech, Université Paris-Saclay, 75013 Paris, France
Email: olivier.rioul@telecom-paristech.fr

***Abstract*—We present a simple proof of the entropy-power inequality using an optimal transportation argument which takes the form of a simple change of variables. The same argument yields a reverse inequality involving a conditional differential entropy which has its own interest. It can also be generalized in various ways. The equality case is easily captured by this method and the proof is formally identical in one and several dimensions.**

I. Introduction

The entropy-power inequality gives a lower bound on the differential entropy of a sum of independent random vectors in terms of their individual differential entropies, and is perhaps the most fascinating inequality that was stated by Shannon in his 1948 seminal paper [15]. To simplify the presentation we assume, without loss of generality, that all considered random vectors have *zero mean* and we first restrict ourselves to real-valued random *variables* in one dimension.

Letting $P(X)=\mathbb{E}\{X^2\}$ be the (average) power of a random variable X, Shannon defined the *entropy-power* $N(X)$ as the *power* of a Gaussian random variable X^* having the same *entropy* as X. He argued [15, § 21] that for continuous random variables it is more convenient to work with the entropy-power $N(X)$ than with the differential entropy $h(X)$.

By the well-known formula $h(X^*)=\frac{1}{2}\log\big(2\pi eP(X^*)\big)$ of the entropy of the Gaussian X^*, the closed-form expression of $N(X)=P(X^*)$ when $h(X^*)=h(X)$ is

$$N(X)=\frac{e^{2h(X)}}{2\pi e} \tag{1}$$

which is essentially e to the *power* twice the *entropy* of X, also the "entropy power" of X in this sense. Since the Gaussian maximizes entropy for a given power: $h(X)\le\frac{1}{2}\log\big(2\pi eP(X)\big)$, the entropy-power does not exceed the actual power: $N(X)\le P(X)$ with equality if and only if X is Gaussian.

A basic property of the entropy-power is the *scaling* property. The power of a scaled random variable is given by $P(aX)=a^2P(X)$, and the same property holds for the entropy-power:

$$N(aX)=a^2N(X) \tag{2}$$

thanks to the well-known scaling property of the entropy:

$$h(aX)=h(X)+\log a \qquad (a>0). \tag{3}$$

For any two *independent* continuous random variables X and Y, the power of the sum equals the sum of the individual powers: $P(X+Y)=P(X)+P(Y)$ and clearly the same relation holds for the entropy-power in the case of Gaussian variables. For non-Gaussian variables, however, the entropy-power of the sum exceeds the sum of the individual entropy-powers:

$$N(X+Y)\ge N(X)+N(Y) \tag{4}$$

where equality holds only if X and Y are Gaussian. This is the celebrated entropy-power inequality (EPI) as stated by Shannon. It is remarkable that Shannon had the intuition of this inequality since it turns out to be quite difficult to prove. The first rigorous proof is due to Stam [16] more than ten years after Shannon's paper and is quite involved.

Thirty years after Shannon's paper, Lieb [9] gave a very different proof of an equivalent entropy-power inequality that is more convenient to prove. By the scaling property (2), one has $N(\sqrt{\lambda}X)=\lambda N(X)$ for any $0<\lambda<1$, and the EPI (4) is clearly equivalent to

$$N(\sqrt{\lambda}X+\sqrt{1-\lambda}Y)\ge\lambda N(X)+(1-\lambda)N(Y). \tag{5}$$

Taking the logarithm on both sides it follows from the concavity of the logarithm that

$$h(\sqrt{\lambda}X+\sqrt{1-\lambda}Y)\ge\lambda h(X)+(1-\lambda)h(Y). \tag{6}$$

Conversely, to prove (5) it is sufficient, by appropriately scaling the variables, to assume that X and Y have the same entropy power $N(X)=N(Y)$, hence the same entropy $h(X)=h(Y)$. In this case, taking the exponential on both sides of (6), the r.h.s. becomes $(e^{2h(X)})^\lambda(e^{2h(Y)})^{1-\lambda}=\lambda e^{2h(X)}+(1-\lambda)e^{2h(Y)}$ which gives (5). Thus Lieb's restatement (6) is equivalent to the EPI. Equality holds in (6) if and only if X and Y are Gaussian with the same power.

Both (5) and (6) have a nice interpretation [5]: both the entropy-power N and the entropy h are *concave* under the *power-preserving* combination $\sqrt{\lambda}X+\sqrt{1-\lambda}Y$. That linear combination is power-preserving because if X and Y have the same power P, then $\sqrt{\lambda}X+\sqrt{1-\lambda}Y$ also has the same power P.

All available proofs of the EPI (6) can be seen as either variants of Stam's proof using a Gaussian perturbation argument (where the entropies are differentiated with respect to the power of an additive Gaussian noise), or variants of Lieb's proof using sharp inequalities from functional analysis such as Young's convolutional inequality (where the EPI is obtained as a limit case). In this paper, we present a new proof from [14] using a *transportation* argument in which the Gaussian distribution is "transported" to another probability distribution by a simple change of variable. The idea is to relate (6) to the case of

equality: let X^*, Y^* be independent Gaussian with the same power, so that

$$h(\sqrt{\lambda}X^* + \sqrt{1-\lambda}Y^*) = \lambda h(X^*) + (1-\lambda)h(Y^*) \quad (7)$$

A *transportation* from X^* to X, and similarly from Y^* to Y, can be made to compare $h(X)$ to $h(X^*)$, $h(Y)$ to $h(Y^*)$, and also $h(\sqrt{\lambda}X + \sqrt{1-\lambda}Y)$ to $h(\sqrt{\lambda}X^* + \sqrt{1-\lambda}Y^*)$. This is described in the following section.

II. Ingredients

Hereafter we assume that the considered random variables have continuous and positive densities. This assumption can be made without loss of generality (see [14] for details). It follows that all considered cumulative distribution functions are continuously differentiable and (strictly) increasing.

The following lemma is the "not Gaussian to Gaussian" lemma 1 used in [11]:

Lemma 1 (Transportation)**.** *There exists an increasing function T such that $T(X^*)$ has the same distribution as X.*

Proof: Let F_X denote the cumulative distribution function of X. Then $\mathbb{P}\{X \le x\} = F_X(x) = F_{X^*}\big(F_{X^*}^{-1}(F_X(x))\big) = \mathbb{P}\{X^* \le F_{X^*}^{-1}(F_X(x)\} = \mathbb{P}\{F_X^{-1}\big(F_{X^*}(X^*)\big) \le x\}$ which proves the lemma with $T = F_X^{-1} \circ F_{X^*}$. ∎

Notice that the lemma is well-known when X^* is uniformly distributed, to justify the inverse transform sampling method.

This function T is sometimes referred to an "optimal transport" [17] because it solves a Monge-Kantorovitch transportation problem of the type:

$$\min_{\substack{(X,X^*)\\ X\sim p_X, X^*\sim p_{X^*}}} \sqrt{\mathbb{E}\{(X-X^*)^2\}}$$

where the marginal densities are fixed and the minimisation of the transportation cost is done on the joint distribution. The resulting minimum is known as the Wasserstein distance $W_2(X, X^*)$. Thus $X = T(X^*)$ is the random variable which is maximally correlated to X^* for fixed marginals; this is a restatement of a well-known Hardy-Littlewood rearrangement inequality and can be generalized to other convex cost functions. This type of optimality was used in [10] to prove Costa's corner point conjecture for the Gaussian interference channel (see also [13]). However, we shall *not* need such an optimality property here.

By Lemma 1, to prove the EPI we can always assume that $X = T(X^*)$ using transport T, and similarly $Y = U(Y^*)$ using another transport U. Thus the EPI can be restated in terms of the Gaussian variables X^*, Y^* as

$$h(\sqrt{\lambda}T(X^*) + \sqrt{1-\lambda}U(Y^*)) \ge \lambda h(T(X^*)) + (1-\lambda)h(U(Y^*)). \quad (8)$$

We have the following well-known lemma (also used in [11]).

Lemma 2 (Change of Variable in the Entropy)**.**

$$h(T(X)) = h(X) + \mathbb{E}\{\log T'(X)\} \quad (9)$$

where $T' > 0$ denotes the derivative of T.

For linear $T(x) = ax$ we recover the scaling property (2). The general proof is similar:

Proof. Make the change of variable $p_{T(X)}(T(x))\,\mathrm{d}T(x) = p_X(x)\,\mathrm{d}x$ in the expression of the entropy $h(T(X)) = -\mathbb{E}\{\log p_{T(X)}(T(X))\} = -\mathbb{E}\{\log(p_X(X)/T'(X))\} = h(X) + \mathbb{E}\{\log T'(X)\}$. □

Lemma 2 allows one to evaluate the differences $h(T(X^*)) - h(X^*)$ and $h(U(Y^*)) - h(Y^*)$. However, the remaining terms $h(\sqrt{\lambda}T(X^*) + \sqrt{1-\lambda}U(Y^*))$ and $h(\sqrt{\lambda}X^* + \sqrt{1-\lambda}Y^*)$ cannot be compared directly because two variables are involved instead of one. However one variable can be fixed by conditioning and an extended version of Lemma 2 can be used:

Lemma 3 (Change of Variable in the Conditional Entropy)**.**

$$h(T_Y(X)|Y) = h(X|Y) + \mathbb{E}\{\log T_Y'(X)\}. \quad (10)$$

Proof. By Lemma 2, we have $h(T_Y(X)|Y = y) = h(X|Y = y) + \mathbb{E}\{\log T_Y'(X)|Y = y\}$ for a fixed value $Y = y$. The result follows by taking the expectation over Y. □

Using these ingredients, a simple proof of the EPI is obtained as shown in the next section.

III. A Simple Proof of the EPI

From Lemma 1 we can assume that $X = T(X^*)$ using transport T and $Y = U(Y^*)$ using transport U. By Lemma 2,

$$\begin{aligned} h(X) &= h(X^*) + \mathbb{E}\{\log T'(X^*)\} \\ h(Y) &= h(Y^*) + \mathbb{E}\{\log U'(Y^*)\}. \end{aligned} \quad (11)$$

It remains to compare $h(\sqrt{\lambda}X + \sqrt{1-\lambda}Y) = h(\sqrt{\lambda}T(X^*) + \sqrt{1-\lambda}U(Y^*))$ to $h(\sqrt{\lambda}X^* + \sqrt{1-\lambda}Y^*)$, which is the entropy if the Gaussian variable $\widetilde{X} = \sqrt{\lambda}\,X^* + \sqrt{1-\lambda}\,Y^*$. Two independent variables are involved in the expression $\sqrt{\lambda}T(X^*) + \sqrt{1-\lambda}U(Y^*)$ which does not depend on $\widetilde{X}$ alone, but rather on the two variables $(\widetilde{X}, \widetilde{Y})$ obtained by rotation from (X^*, Y^*):

$$\begin{pmatrix} \widetilde{X} \\ \widetilde{Y} \end{pmatrix} = \begin{pmatrix} \sqrt{\lambda} & \sqrt{1-\lambda} \\ -\sqrt{1-\lambda} & \sqrt{\lambda} \end{pmatrix} \begin{pmatrix} X^* \\ Y^* \end{pmatrix}. \quad (12)$$

The inverse rotation reads

$$\begin{pmatrix} X^* \\ Y^* \end{pmatrix} = \begin{pmatrix} \sqrt{\lambda} & -\sqrt{1-\lambda} \\ \sqrt{1-\lambda} & \sqrt{\lambda} \end{pmatrix} \begin{pmatrix} \widetilde{X} \\ \widetilde{Y} \end{pmatrix} \quad (13)$$

which gives $\sqrt{\lambda}T(X^*) + \sqrt{1-\lambda}U(Y^*) = \sqrt{\lambda}T(\sqrt{\lambda}\widetilde{X} - \sqrt{1-\lambda}\widetilde{Y}) + \sqrt{1-\lambda}U(\sqrt{1-\lambda}\widetilde{X} + \sqrt{\lambda}\widetilde{Y})$, a function of $(\widetilde{X}, \widetilde{Y})$ which we denote by $T_{\widetilde{Y}}(\widetilde{X})$. Now since conditioning reduces entropy,

$$h(\sqrt{\lambda}X + \sqrt{1-\lambda}Y) = h(T_{\widetilde{Y}}(\widetilde{X})) \ge h(T_{\widetilde{Y}}(\widetilde{X})|\widetilde{Y}). \quad (14)$$

Lemma 3 applies with

$$\begin{aligned} T_{\widetilde{Y}}'(\widetilde{X}) &= \lambda T'(\sqrt{\lambda}\widetilde{X} - \sqrt{1-\lambda}\widetilde{Y}) + (1-\lambda)U'(\sqrt{1-\lambda}\widetilde{X} + \sqrt{\lambda}\widetilde{Y}) \\ &= \lambda T'(X^*) + (1-\lambda)U'(Y^*) \end{aligned} \quad (15)$$

which gives

$$h(T_{\widetilde{Y}}(\widetilde{X})|\widetilde{Y}) = h(\widetilde{X}|\widetilde{Y}) + \mathbb{E}\{\log\big(\lambda T'(X^*) + (1-\lambda)U'(Y^*)\big)\}. \tag{16}$$

Since X^*, Y^* are independent Gaussian with identical powers, so are the rotated variables $\widetilde{X}, \widetilde{Y}$. By independence,

$$h(\widetilde{X}|\widetilde{Y}) = h(\widetilde{X}) = h(\sqrt{\lambda}\, X^* + \sqrt{1-\lambda}\, Y^*). \tag{17}$$

Therefore, combining (14), (16) and (17) we obtain

$$\begin{aligned} h(\sqrt{\lambda}X + \sqrt{1-\lambda}Y) &\geq h(\sqrt{\lambda}\, X^* + \sqrt{1-\lambda}\, Y^*) \\ &\quad + \mathbb{E}\{\log\big(\lambda T'(X^*) + (1-\lambda)U'(Y^*)\big)\}. \end{aligned} \tag{18}$$

With (11) we conclude that

$$\begin{aligned} &h(\sqrt{\lambda}X + \sqrt{1-\lambda}Y) - \lambda h(X) - (1-\lambda)h(Y) \\ &\geq h(\sqrt{\lambda}X^* + \sqrt{1-\lambda}Y^*) - \lambda h(X^*) - (1-\lambda)h(Y^*) \\ &\quad + \mathbb{E}\{\log\big(\lambda T'(X^*) + (1-\lambda)U'(Y^*)\big)\} \\ &\quad - \lambda\mathbb{E}\{\log T'(X^*)\} - (1-\lambda)\mathbb{E}\{\log U'(Y^*)\} \end{aligned} \tag{19}$$

where the first line in the r.h.s. vanishes by (7) and the remaining part is ≥ 0 by Jensen's inequality (concavity of the logarithm). This proves the EPI (6). □

IV. The Equality Case

The equality case is easily captured by the above method. If equality holds in (19) then

$$\begin{aligned} &\log\big(\lambda T'(X^*) + (1-\lambda)U'(Y^*)\big) \\ &\quad = \lambda\mathbb{E}\{\log T'(X^*)\} + (1-\lambda)\mathbb{E}\{\log U'(Y^*)\} \text{ a.e.} \end{aligned} \tag{20}$$

Because the logarithm is strictly concave and $0 < \lambda < 1$, this implies

$$T'(X^*) = U'(Y^*) \text{ a.e.} \tag{21}$$

Since X^* and Y^* are independent, it follows that T' and U' are constant and equal, hence T and U are linear and $X = c \cdot X^*$, $Y = c \cdot Y^*$ are Gaussian with the same power. This is the required equality case of the EPI (6). Of course this condition also implies equality in (14) since then $\sqrt{\lambda}X + \sqrt{1-\lambda}Y = T_{\widetilde{Y}}(\widetilde{X}) = c\widetilde{X}$ is independent of $\widetilde{Y}$.

V. Generalization to random vectors

The above proof of the EPI carries over *verbatim* to random vectors in n dimensions. The only change is that transport maps $T : \mathbb{R}^n \to \mathbb{R}^n$ are n-dimensional—accordingly, T' denotes the *Jacobian* determinant of T. Lemma 1 is easily extended to random vectors using the so-called *Knöthe's map* in the theory of convex bodies [8], [17], of the form

$$T(x) = \big(T_1(x_1), T_2(x_1, x_2), \ldots, T_n(x_1, \ldots, x_n)\big) \tag{22}$$

where $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$. The Jacobian matrix of T is triangular with positive diagonal elements:

$$\begin{pmatrix} \frac{\partial T_1}{\partial x_1} & 0 & \cdots & 0 \\ \frac{\partial T_2}{\partial x_1} & \frac{\partial T_2}{\partial x_2} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial T_n}{\partial x_1} & \frac{\partial T_n}{\partial x_2} & \cdots & \frac{\partial T_n}{\partial x_n} \end{pmatrix} \tag{23}$$

so that

$$T'(x) = \prod_{i=1}^{n} \frac{\partial T_i}{\partial x_i} > 0. \tag{24}$$

This transport map was used in [11] and details about its construction can also be found in [14]. Lemmas 2 and 3 are then obtained by a change of variable in n dimensions. The above proof of the EPI is identical word for word, where the concavity of the logarithm in the last step (19) is used on each dimension.

VI. A Reverse EPI

A. Derivation: Generalization to non-Gaussian X^* and Y^*

The above proof of the EPI can also be generalized to the case where X^* and Y^* are *not* necessarily Gaussian. In fact a closer look at the proof reveals that the Gaussian assumption is never used except for the simplification in (17) which relies on the independence of $\widetilde{X} = \sqrt{\lambda}X^* + \sqrt{1-\lambda}Y^*$ and $\widetilde{Y} = -\sqrt{1-\lambda}X^* + \sqrt{\lambda}Y^*$. If such an independence does not hold, we obtain the more general inequality

$$\begin{aligned} &h(\sqrt{\lambda}X + \sqrt{1-\lambda}Y) - \lambda h(X) - (1-\lambda)h(Y) \\ &\geq h(\sqrt{\lambda}X^* + \sqrt{1-\lambda}Y^* | -\sqrt{1-\lambda}X^* + \sqrt{\lambda}Y^*) \\ &\quad - \lambda h(X^*) - (1-\lambda)h(Y^*) \end{aligned} \tag{25}$$

valid for any independent X, Y and any independent X^*, Y^*. In fact this gives two independent inequalities: For Gaussian X^*, Y^* the r.h.s. vanishes and we recover the classical EPI. But for Gaussian X, Y the l.h.s. vanishes, so that the r.h.s. is ≤ 0, and we obtain a *reverse* inequality which (rewritten for X, Y) takes the form

$$h(\sqrt{\lambda}X + \sqrt{1-\lambda}Y | -\sqrt{1-\lambda}X + \sqrt{\lambda}Y) \leq \lambda h(X) + (1-\lambda)h(Y). \tag{26}$$

Compared to (6), the opposite inequality holds but for a conditional differential entropy. In other words, $\lambda h(X) + (1-\lambda)h(Y)$ is upper bounded by the differential entropy of $\sqrt{\lambda}X + \sqrt{1-\lambda}Y$ and lower bounded by its conditional differential entropy given $-\sqrt{1-\lambda}X + \sqrt{\lambda}Y$, the difference between the bounds being equal to the mutual information $I(\sqrt{\lambda}X + \sqrt{1-\lambda}Y; -\sqrt{1-\lambda}X + \sqrt{\lambda}Y)$. Thus an equivalent restatement is

$$\begin{aligned} 0 &\leq h(\sqrt{\lambda}\, X + \sqrt{1-\lambda}\, Y) - \lambda h(X) - (1-\lambda)h(Y) \\ &\leq I(\sqrt{\lambda}\, X + \sqrt{1-\lambda}\, Y; -\sqrt{1-\lambda}\, X + \sqrt{\lambda}\, Y). \end{aligned} \tag{27}$$

This mutual information can be seen as an upper bound on the *deficit* in the EPI for X and Y, which is zero if and only if X and Y are Gaussian with identical powers. Courtade [3] recently derived a similar bound on the deficit in the logarithmic Sobolev inequality, which is equivalent to another type of "reverse EPI". As above the extension to random vectors in n dimensions is straightforward.

$$
\begin{aligned}
\mathbf{K}_{U|V} &= \lambda\mathbf{K}_X + (1-\lambda)\mathbf{K}_Y - \lambda(1-\lambda)(\mathbf{K}_Y - \mathbf{K}_X)\big[(1-\lambda)\mathbf{K}_X + \lambda\mathbf{K}_Y\big]^{-1}(\mathbf{K}_Y - \mathbf{K}_X) \\
&= \lambda\mathbf{K}_X - \lambda(1-\lambda)\mathbf{K}_X\big[(1-\lambda)\mathbf{K}_X + \lambda\mathbf{K}_Y\big]^{-1}\mathbf{K}_X + \lambda(1-\lambda)\mathbf{K}_X\big[(1-\lambda)\mathbf{K}_X + \lambda\mathbf{K}_Y\big]^{-1}\mathbf{K}_Y \\
&\quad + (1-\lambda)\mathbf{K}_Y - \lambda(1-\lambda)\mathbf{K}_Y\big[(1-\lambda)\mathbf{K}_X + \lambda\mathbf{K}_Y\big]^{-1}\mathbf{K}_Y + \lambda(1-\lambda)\mathbf{K}_Y\big[(1-\lambda)\mathbf{K}_X + \lambda\mathbf{K}_Y\big]^{-1}\mathbf{K}_X \\
&= \lambda\mathbf{K}_X\big[(1-\lambda)\mathbf{K}_X + \lambda\mathbf{K}_Y\big]^{-1}\big((1-\lambda)\mathbf{K}_X + \lambda\mathbf{K}_Y - (1-\lambda)\mathbf{K}_X\big) + \lambda(1-\lambda)\mathbf{K}_X\big[(1-\lambda)\mathbf{K}_X + \lambda\mathbf{K}_Y\big]^{-1}\mathbf{K}_Y \\
&\quad + (1-\lambda)\mathbf{K}_Y\big[(1-\lambda)\mathbf{K}_X + \lambda\mathbf{K}_Y\big]^{-1}\big((1-\lambda)\mathbf{K}_X + \lambda\mathbf{K}_Y - \lambda\mathbf{K}_Y\big) + \lambda(1-\lambda)\mathbf{K}_Y\big[(1-\lambda)\mathbf{K}_X + \lambda\mathbf{K}_Y\big]^{-1}\mathbf{K}_X \\
&= \big(\lambda^2 + \lambda(1-\lambda) + (1-\lambda)^2 + \lambda(1-\lambda)\big)\big[\lambda\mathbf{K}_X^{-1} + (1-\lambda)\mathbf{K}_Y^{-1}\big]^{-1} = \big[\lambda\mathbf{K}_X^{-1} + (1-\lambda)\mathbf{K}_Y^{-1}\big]^{-1}.
\end{aligned}
$$

B. The Equality Case and Bernstein's Lemma

We have seen that equality holds in the above proof of the EPI if and only if (X, Y) and (X^*, Y^*) are proportional. The same argument shows that the the same equality condition holds for the reverse EPI. Thus both the EPI (6) and its reverse (26) are equalities if and only if X and Y are i.i.d. Gaussian. This also corresponds to the case where the mutual information vanishes in (27). This gives an alternative proof of Bernstein's lemma (see e.g., [6, Appendix I] and [2, Chap. 5]):

Lemma 4 (Bernstein)**.** *Let X and Y be independent. Then the rotated $\sqrt{\lambda}\,X + \sqrt{1-\lambda}\,Y$, $-\sqrt{1-\lambda}\,X + \sqrt{\lambda}\,Y$ are independent if and only if X, Y are i.i.d. Gaussian.*

C. The Gaussian Case

If X and Y are Gaussian with not necessarily equal powers $P(X)$ and $P(Y)$, it is easily seen that (26) and (6) reduce to the harmonic/geometric/arithmetic inequalities

$$
\begin{aligned}
\big(\lambda P(X)^{-1} + (1-\lambda)P(Y)^{-1}\big)^{-1} &\\
&\le P(X)^{\lambda}P(Y)^{1-\lambda} \\
&\le \lambda P(X) + (1-\lambda)P(Y). \quad (28)
\end{aligned}
$$

More generally for Gaussian vectors, if $X \sim \mathcal{N}(0, \mathbf{K}_X)$ and $Y \sim \mathcal{N}(0, \mathbf{K}_Y)$ not necessarily of identical covariances, it is known [5, Thm. 8] that the EPI reduces to Ky Fan's concavity inequality of the log-determinant: using the well-known formula $h(U) = \frac{1}{2}\log\big((2\pi e)^n|\mathbf{K}_U|\big)$ the EPI reduces to $\log|\lambda\mathbf{K}_X + (1-\lambda)\mathbf{K}_Y| \ge \lambda\log|\mathbf{K}_X| + (1-\lambda)\log|\mathbf{K}_Y|$.

Similarly for the reverse EPI, noting that $h(U|V) = \frac{1}{2}\log\big((2\pi e)^n|\mathbf{K}_{U|V}|\big)$ where $\mathbf{K}_{U|V}$ is Schur's complement $\mathbf{K}_{U|V} = \mathbf{K}_U - \mathbf{K}_{UV}\mathbf{K}_V^{-1}\mathbf{K}_{VU}$ (where $\mathbf{K}_{UV}$ is an inter-covariance matrix), set $U = \sqrt{\lambda}\,X + \sqrt{1-\lambda}\,Y$ and $V = -\sqrt{1-\lambda}\,X + \sqrt{\lambda}\,Y$, $\mathbf{K}_U = \lambda\mathbf{K}_X + (1-\lambda)\mathbf{K}_Y$, $\mathbf{K}_V = (1-\lambda)\mathbf{K}_X + \lambda\mathbf{K}_Y$, and $\mathbf{K}_{UV} = \mathbf{K}_{VU} = \sqrt{\lambda(1-\lambda)}(\mathbf{K}_Y - \mathbf{K}_X)$. By the calculation shown at the top of this page, the reverse EPI reduces to the inequality $\log|\lambda\mathbf{K}_X^{-1} + (1-\lambda)\mathbf{K}_Y^{-1}|^{-1} \le \lambda\log|\mathbf{K}_X| + (1-\lambda)\log|\mathbf{K}_Y|$. Thus (26) and (6) reduce to the generalized harmonic/geometric/arithmetic inequalities:

$$
\begin{aligned}
|\lambda\mathbf{K}_X^{-1} + (1-\lambda)\mathbf{K}_Y^{-1}|^{-1} &\\
&\le |\mathbf{K}_X|^{\lambda}|\mathbf{K}_Y|^{1-\lambda} \\
&\le |\lambda\mathbf{K}_X + (1-\lambda)\mathbf{K}_Y|. \quad (29)
\end{aligned}
$$

D. Equivalence Between the EPI and its Reverse

As observed by Chandra Nair in a private communication to the author, it turns out that the reverse EPI is in fact equivalent to the EPI where the roles of X and Y are permuted. In fact (26) is equivalent to

$$
\begin{aligned}
h(\sqrt{\lambda}\,X + \sqrt{1-\lambda}\,Y, -\sqrt{1-\lambda}\,X + \sqrt{\lambda}\,Y) &\\
&\le \lambda h(X) + (1-\lambda)h(Y) + h(-\sqrt{1-\lambda}\,X + \sqrt{\lambda}\,Y) \quad (30)
\end{aligned}
$$

where the joint entropy in the r.h.s. equals $h(X, Y) = h(X) + h(Y)$ by the scaling property of the differential entropy for vectors. Reorganizing terms one obtains the following version of the EPI:

$$
(1-\lambda)h(X) + \lambda h(Y) \le h(-\sqrt{1-\lambda}\,X + \sqrt{\lambda}\,Y). \quad (31)
$$

We recover, in particular, that the cases of equality are the same for the EPI and its reverse. The above calculation was already used by Wang and Madiman [18] as a short proof of the EPI under the hypothesis that X and Y follow symmetrical and identical distributions. One reason why the EPI is equivalent to its reverse version is suggested below in relation to Young's convolutional inequality and its reverse.

VII. Zamir and Feder's Generalization to Linear Transformations

An immediate generalization of the EPI (6) for n independent variables $X_1, X_2, \ldots, X_n$ is

$$
h\big(\sum_i a_i X_i\big) \ge \sum_i a_i^2 h(X_i) \quad (32)
$$

where the coefficients are normalized such that $\sum_i a_i^2 = 1$. The above proof of the EPI can easily be adapted to prove this inequality directly by letting $\mathbf{A}$ be an orthogonal matrix whose first line is $(a_1, a_2, \ldots, a_n)$ and defining

$$
\widetilde{X} = \mathbf{A}X^* \quad (33)
$$

where $\widetilde{X^*}$ is a column vector of n i.i.d. Gaussian variables $X_1^*, X_2^*, \ldots, X_n^*$. The inverse transformation is $X^* = \mathbf{A}^t\widetilde{X}$ and the proof is easily modified along these lines.

Essentially the same proof can be used for Zamir and Feder's generalized EPI [19] (see also [12, § IV]):

$$
h(\mathbf{A}X) \ge \sum_{i,j} a_{i,j}^2 h(X_j) \quad (34)
$$

where X is the column vector of components $X_1, X_2, \ldots, X_n$ and $\mathbf{A} = (a_{i,j})$ is any real-valued (possibly rectangular) matrix with orthonormal rows. By adding orthonormal rows we form a square orthonormal matrix (still denoted by $\mathbf{A}$) and the same transformation $\widetilde{X} = \mathbf{A}X^*$, $X^* = \mathbf{A}^t\widetilde{X}$ is used. The conclusion follows from a simple inequality [7, Lemma 1] which generalizes Jensen's inequality for the logarithm.

VIII. Generalization to Rényi Entropies and the Relation to Young's Inequality

The Rényi entropy of order $p > 0$ ($p \neq 1$) is defined as

$$h_p(X) = h_p(f) = \frac{1}{1-p}\log\int f^p = -p'\log\|f\|_p \qquad (35)$$

where $\|f\|_p$ denotes the L^p norm of the density f of X and p' is p's conjugate such that $1/p + 1/p' = 1$. While the above proof of the EPI focuses on the equivalent inequality

$$\begin{aligned} &h(\sqrt{\lambda}X + \sqrt{1-\lambda}\,Y) - \lambda h(X) - (1-\lambda)h(Y) \\ &\quad \geq h(\sqrt{\lambda}X^* + \sqrt{1-\lambda}\,Y^*) - \lambda h(X^*) - (1-\lambda)h(Y^*) \end{aligned} \qquad (36)$$

for i.i.d. Gaussian X^*, Y^*, the natural generalization of the EPI considered in [5] takes the form

$$\begin{aligned} &h_r(\sqrt{\lambda}X + \sqrt{1-\lambda}\,Y) - \lambda h_p(X) - (1-\lambda)h_q(Y) \\ &\quad \geq h_r(\sqrt{\lambda}X^* + \sqrt{1-\lambda}\,Y^*) - \lambda h_p(X^*) - (1-\lambda)h_q(Y^*) \end{aligned} \qquad (37)$$

where p, q, r are chosen such that $1/p' = \lambda/r'$ and $1/q' = (1-\lambda)/r'$ (so that $1/p' + 1/q' = 1/r'$), where p', q', and r' are the conjugates of p, q and r, respectively.

The EPI can then be obtained by letting $p, q, r \to 1$ as shown in [5]. This is the preferred proof of the EPI in the classical textbook by Cover and Thomas [4]. Notice that because the EPI is obtained as a limit, the equality case is not settled by this method.

The above transportation proof can also be generalized, with the same transport maps $X = T(X^*)$ and $Y = U(Y^*)$, to prove (37). Now the Rényi EPI (37) is in fact equivalent to the sharp Young's convolutional inequality or its reverse inequality [4], [5]. This is easily seen by noting that the Rényi entropy enjoys the same scaling property (3) as the differential entropy so that (37) is equivalent to saying that $h_r(\sqrt{\lambda}X + \sqrt{1-\lambda}\,Y) - \lambda h_p(\sqrt{\lambda}X) - (1-\lambda)h_q(\sqrt{1-\lambda}\,Y)$ (which equals $-r'\log\|f * g\|_r + r'\log\|f\|_p + r'\log\|g\|_q$ where f and g denote the densities of $\sqrt{\lambda}X$ and $\sqrt{1-\lambda}Y$) is minimized for i.i.d. Gaussian X and Y. Dividing by r' gives sharp Young's inequality for $p, q, r > 1$ ($r' > 0$):

$$\sqrt{\frac{r^{1/r}}{r'^{1/r'}}}\|f * g\|_r \leq \sqrt{\frac{p^{1/p}}{p'^{1/p'}}}\|f\|_p \cdot \sqrt{\frac{q^{1/q}}{q'^{1/q'}}}\|g\|_q \qquad (38)$$

and the *reverse* Young's inequality for $0 < p, q, r < 1$ ($r' < 0$):

$$\sqrt{\frac{r^{1/r}}{|r'|^{1/r'}}}\|f * g\|_r \geq \sqrt{\frac{p^{1/p}}{|p'|^{1/p'}}}\|f\|_p \cdot \sqrt{\frac{q^{1/q}}{|q'|^{1/q'}}}\|g\|_q. \qquad (39)$$

In fact Barthe [1] gave a transportation proof of both inequalities. Since one obtains the EPI by letting $p, q, r \to 1^+$ from above (from Young's inequality) and also by letting $p, q, r \to 1^-$ from below (from the reverse Young's inequality), the EPI and its reverse are equivalent at the limit $p, q, r \to 1$.

Acknowledgment

The author would like to thank Tom Courtade, Chandra Nair and Michèle Wigger for their discussions.

References

[1] F. Barthe, "Optimal Young's inequality and its converse: A simple proof," *GAFA, Geom. funct. anal.*, vol. 8, no. 2, pp. 234–242, 1998.

[2] W. Bryc, *The Normal Distribution - Characterizations with Applications*, ser. Lecture Notes in Statistics. Springer, 1995, vol. 100.

[3] T. A. Courtade, "Links between the logarithmic Sobolev inequality and the convolution inequalities for entropy and Fisher information," 2017, draft available at https://arxiv.org/abs/1608.05431.

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

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

[6] Y. Geng and C. Nair, "The capacity region of the two-receiver Gaussian vector broadcast channel with private and common messages," *IEEE Trans. Inf. Theory*, vol. 60, no. 4, pp. 2087–2104, Apr. 2014.

[7] D. Guo, S. Shamai (Shitz), and S. Verdú, "Proof of entropy power inequalities via MMSE," in *Proc. IEEE Int. Symp. Information Theory*, Seattle, USA, Jul. 2006, pp. 1011–1015.

[8] H. Knothe, "Contributions to the theory of convex bodies," *Michigan Math. J.*, vol. 4, pp. 39–52, 1957.

[9] E. H. Lieb, "Proof of an entropy conjecture of Wehrl," *Commun. Math. Phys.*, vol. 62, pp. 35–41, 1978.

[10] Y. Polyanskiy and Y. Wu, "Wasserstein continuity of entropy and outer bounds for interference channels," *IEEE Trans. Inf. Theory*, vol. 62, no. 7, pp. 3992–4002, Jul. 2016.

[11] O. Rioul and M. H. M. Costa, "On some almost properties," in *IEEE Information Theory and Applications Workshop (ITA 2016)*, San Diego, USA, Feb. 2016.

[12] O. Rioul, "Information theoretic proofs of entropy power inequalities," *IEEE Trans. Inf. Theory*, vol. 57, no. 1, pp. 33–55, Jan. 2011.

[13] ——, "At every corner: Determining corner points of two-user Gaussian interference channels," in *IEEE International Conference on Communications (ICC'17)*, Paris, France, May 21–25, 2017.

[14] ——, "Yet another proof of the entropy power inequality," *IEEE Trans. Inf. Theory*, to appear., draft available at https://arxiv.org/abs/1606.05969.

[15] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 623–656, Oct. 1948.

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

[17] C. Villani, *Optimal Transport: Old and New*, ser. Science and Business Media. Springer, 2008, vol. 338.

[18] L. Wang and M. Madiman, "A new approach to the entropy power inequality, via rearrangements," in *Proc. IEEE Int. Symp. Information Theory*, Istanbul, Turkey, Jul. 2013, pp. 599–603.

[19] R. Zamir and M. Feder, "A generalization of information theoretic inequalities to linear transformations of independent vector," in *Proceedings of the 6-th Joint Swedish-Russian International Workshop on Information Theory*, Möle, Sweden, Aug. 1993, pp. 254–258.