Yet Another Proof of the Entropy Power Inequality

Olivier Rioul, *Member, IEEE*

***Abstract*—Yet another simple proof of the entropy power inequality is given, which avoids both the integration over a path of Gaussian perturbation and the use of Young's inequality with sharp constant or Rényi entropies. The proof is based on a simple change of variables, is formally identical in one and several dimensions, and easily settles the equality case.**

***Index Terms*—Entropy power inequality, differential entropy, gaussian variables, optimal transport.**

I. Introduction

THE entropy power inequality (EPI) was stated by Shannon [1] in the form

$$e^{\frac{2}{n}h(X+Y)} \geq e^{\frac{2}{n}h(X)} + e^{\frac{2}{n}h(Y)} \tag{1}$$

for any independent n-dimensional random vectors $X, Y \in \mathbb{R}^n$ with densities and finite second moments, with equality if and only if X and Y are Gaussian with proportional covariances. Shannon gave an incomplete proof; the first complete proof was given by Stam [2] using properties of Fisher's information. A detailed version of Stam's proof was given by Blachman [3]. A very different proof was provided by Lieb [4] using Young's convolutional inequality with sharp constant. Dembo *et al.* [5] provided a clear exposition of both Stam's and Lieb's proofs. Carlen and Soffer gave an interesting variation of Stam's proof for one-dimensional variables [6]. Szarek and Voiculescu [7] gave a proof related to Lieb's but based on a variant of the Brunn-Minkowski inequality. Guo *et al.* [8], Verdú and Guo [9] gave another proof based on the I-MMSE relation. A similar proof based on a relation between divergence and causal MMSE was given by Binia [10]. Yet another proof based on properties of mutual information was proposed in [11] and [12]. A more involved proof based on a stronger form of the EPI that uses spherically symmetric rearrangements, also related to Young's inequality with sharp constant, was recently given by Wang and Madiman [13].

As first noted by Lieb [4], the above Shannon's formulation (1) of the EPI is equivalent to

$$h(\sqrt{\lambda}X + \sqrt{1-\lambda}\,Y) \geq \lambda h(X) + (1-\lambda)h(Y) \tag{2}$$

for any $0 < \lambda < 1$. All available proofs of the EPI used this form.[1] Proofs of the equivalence can be found in numerous papers, e.g., [5, Ths. 4, 6, and 7], [9, Lemma 1], [12, Prop. 2], and [14, Th. 2.5].

There are a few technical difficulties for proving (2) which are not always explicitly stated in previous proofs. First of all, one should check that for any random vector X with finite second moments, the differential entropy $h(X)$ is always well-defined—even though it could be equal to $-\infty$. This is a consequence of [12, Proposition 1]; see also Appendix A for a precise statement and proof. Now if both independent random vectors X and Y have densities and finite second moments, so has $\sqrt{\lambda}X+\sqrt{1-\lambda}\,Y$ and both sides of (2) are well-defined. Moreover, if either $h(X)$ or $h(Y)$ equals $-\infty$ then (2) is obviously satisfied. Therefore, one can always assume that X and Y have *finite* differential entropies.[2]

Another technical difficulty is the requirement for smooth densities. More precisely, as noted in [13, Remark 10] some previous proofs use implicitly that for any X with arbitrary density and finite second moments and any Gaussian[3] Z independent of X,

$$\lim_{t\downarrow 0} h(X+\sqrt{t}Z) = h(X). \tag{3}$$

This was proved explicitly in [12, Lemma 3] and [13, Th. 6.2] using the lower-semicontinuity of divergence; the same result can also be found in previous works that were not directly related to the EPI [16, eq. (51)], [17, Proof of Lemma 1], [18, Proof of Th. 1].

As a consequence, it is sufficient to prove the EPI for random vectors of the form $X+\sqrt{t}Z$ ($t>0$). Indeed, letting Z' be an independent copy of Z such that (Z,Z') is independent of (X,Y), the EPI written for $X+\sqrt{t}Z$ and $Y+\sqrt{t}Z'$ reads

$$\begin{aligned} &h(\sqrt{\lambda}X + \sqrt{1-\lambda}\,Y + \sqrt{t}Z'') \\ &\qquad \geq \lambda h(X+\sqrt{t}Z) + (1-\lambda)h(Y+\sqrt{t}Z') \end{aligned}$$

where $Z'' = \sqrt{\lambda}Z + \sqrt{1-\lambda}\,Z'$ is again identically distributed as Z and Z'. Letting $t\to 0$ we obtain the general EPI (2).[4] Now, for any random vector X and any $t>0$, $X+\sqrt{t}Z$ has a continuous and positive density. This can be seen using the properties of the characteristic function, similarly as in [12, Lemma 1]; see Appendix B for a precise statement and proof. Therefore, as already noticed in [13, Sec. XI], one can always assume that X and Y have *continuous, positive* densities.

One is thus led to prove the following version of the EPI.

Manuscript received June 21, 2016; revised January 23, 2017; accepted February 18, 2017. Date of publication March 1, 2017; date of current version May 18, 2017.

The author is with the LTCI, Telecom ParisTech, Paris-Saclay University, 75013 Paris, France (e-mail: olivier.rioul@telecom-paristech.fr).

[1] Stam's original proof [2] is an exception, but it was later simplified by Dembo, Cover and Thomas [5] using this form.

[2] A nice discussion of general necessary and sufficient conditions for the EPI (1) can be found in [15, Secs. V and VI].

[3] Throughout this paper we assume that Gaussian random vectors are non-degenerate (have non-singular covariance matrices).

[4] A similar observation was done in [6] in a different context of the Ornstein-Uhlenbeck semigroup (instead of the heat semigroup).

Theorem (EPI): Let X, Y be independent random vectors with continuous, positive densities and finite differential entropies and second moments. For any $0 < \lambda < 1$,

$$h(\sqrt{\lambda}X + \sqrt{1-\lambda}\,Y) \geq \lambda h(X) + (1-\lambda)h(Y) \tag{2}$$

with equality if and only if X, Y are Gaussian with identical covariances.

Previous proofs of (2) can be classified into two categories:

- proofs in [2], [3], [6], and [8]–[12] rely on the integration over a path of a continuous Gaussian perturbation of some data processing inequality using either Fisher's information, the minimum mean-squared error (MMSE) or mutual information. As explained in [11, eq. (10)], [12] and [19, eq. (25)], it is interesting to note that in this context, Fisher's information and MMSE are complementary quantities;
- proofs in [4], [7], [13], and [20] are related to Young's inequality with sharp constant or to an equivalent argumentation using spherically symmetric rearrangements, and/or the consideration of convergence of Rényi entropies.

It should also be noted that not all of the available proofs of (2) settle the *equality* case—that equality in (2) holds *only* for Gaussian random vectors with identical covariances. Only proofs from the first category using Fisher's information were shown to capture the equality case. This was made explicit by Stam [2], Carlen and Soffer [6] and for more general fractional EPI's by Madiman and Barron [19].

In this paper, a simple proof of the Theorem is given that avoids both the integration over a path of a continuous Gaussian perturbation and the use of Young's inequality, spherically symmetric rearrangements, or Rényi entropies. It is based on a "Gaussian to not Gaussian" lemma proposed in [21] and is formally identical in one dimension ($n = 1$) and in several dimensions ($n > 1$). It also easily settles the equality case.

II. From Gaussian to Not Gaussian

The following "Gaussian to not Gaussian" lemma [21] will be used here only in the case where X^* is a n-variate Gaussian vector, e.g., $X^* \sim \mathcal{N}(0, \mathbf{I})$, but holds more generally as X^* needs not be Gaussian.

Lemma 1: Let $X = (X_1, \ldots, X_n)$ and $X^* = (X_1^*, \ldots, X_n^*)$ be any two n-dimensional random vectors in $\mathbb{R}^n$ with continuous, positive densities. There exists a diffeomorphism Φ whose Jacobian matrix is triangular with positive diagonal elements such that X has the same distribution as $\Phi(X^*)$.

For completeness we present two proofs in the Appendix. The first proof in Appendix C follows Knöthe [22]. The second proof in Appendix D is based on the (multivariate) inverse sampling method.

The essential content of this lemma is well known in the theory of convex bodies [23, p. 126], [24, Th. 3.4], [25, Th. 1.3.1] where Φ is known as the *Knöthe map* between two convex bodies. The difference with Knöthe's map is that in Lemma 1, the determinant of the Jacobian matrix need not be constant. The Knöthe map is also closely related to the so-called *Knöthe-Rosenblatt coupling* in optimal transport theory [26], [27], and there is a large literature of optimal transportation arguments for geometric-functional inequalities such as the Brunn-Minkowki, isoperimetric, sharp Young, sharp Sobolev and Prékopa-Leindler inequalities. The Knöthe map was used in the original paper by Knöthe [22] to generalize the Brunn-Minkowski inequality, by Gromov in [23, Appendix I] to obtain isoperimetric inequalities on manifolds and by Barthe [28] to prove the sharp Young's inequality. In a similar vein, other transport maps such as the Brenier map were used in [29] for sharp Sobolev and Gagliardo-Nirenberg inequalities and in [30] for a generalized Prékopa-Leindler inequality on manifolds with lower Ricci curvature bounds. Since the present paper was submitted, the Brenier map has also been applied to the stability of the EPI for log-concave densities [31]. All the above-mentionned geometric-functional inequalities are known to be closely related to the EPI (see e.g., [5]), and it is perhaps not too surprising to expect a direct proof of the EPI using an optimal transportation argument—namely, Knöthe map—which is what this paper is about.

Let Φ' be the *Jacobian* (i.e., the determinant of the Jacobian matrix) of Φ. Since $\Phi' > 0$, the usual change of variable formula reads

$$\int f(x)\,\mathrm{d}x = \int f(\Phi(x^*))\Phi'(x^*)\,\mathrm{d}x^*. \tag{4}$$

A simple application of this formula gives the following well-known lemma which was used in [21].

Lemma 2: For any diffeomorphism Φ with positive Jacobian $\Phi' > 0$, if $h\big(\Phi(X^*)\big)$ is finite,

$$h\big(\Phi(X^*)\big) = h(X^*) + \mathbb{E}\{\log \Phi'(X^*)\}. \tag{5}$$

The proof is given for completeness.

Proof: Let $f(x)$ be the density of $\Phi(X^*)$ so that $g(x^*) = f(\Phi(x^*))\Phi'(x^*)$ is the density of X^*. Then we have $\int f(x)\log f(x)\,\mathrm{d}x = \int f(\Phi(x^*))\log f(\Phi(x^*)) \cdot \Phi'(x^*)\mathrm{d}x^* = \int g(x^*)\log\big(g(x^*)/\Phi'(x^*)\big)\,\mathrm{d}x^*$ which yields (5). □

III. Proof of the Entropy Power Inequality

Let X^*, Y^* be *any* i.i.d. Gaussian random vectors, e.g., $\sim\mathcal{N}(0, \mathbf{I})$. For any $0 < \lambda < 1$, $\sqrt{\lambda}X^* + \sqrt{1-\lambda}\,Y^*$ is identically distributed as X^* and Y^* and, therefore,

$$h(\sqrt{\lambda}X^* + \sqrt{1-\lambda}\,Y^*) = \lambda h(X^*) + (1-\lambda)h(Y^*). \tag{6}$$

Subtracting both sides from both sides of (2) one is led to prove that

$$\begin{aligned} &h(\sqrt{\lambda}X + \sqrt{1-\lambda}\,Y) - h(\sqrt{\lambda}X^* + \sqrt{1-\lambda}\,Y^*) \\ &\qquad \geq \lambda\big(h(X) - h(X^*)\big) + (1-\lambda)\big(h(Y) - h(Y^*)\big). \end{aligned} \tag{7}$$

Let Φ be as in Lemma 1, so that X has the same distribution as $\Phi(X^*)$. Similarly let Ψ be such that Y has the same distribution as $\Psi(Y^*)$. Since $\sqrt{\lambda}X + \sqrt{1-\lambda}\,Y$ is identically distributed as $\sqrt{\lambda}\Phi(X^*) + \sqrt{1-\lambda}\,\Psi(Y^*)$,

$$\begin{aligned} &h(\sqrt{\lambda}X + \sqrt{1-\lambda}\,Y) - h(\sqrt{\lambda}X^* + \sqrt{1-\lambda}\,Y^*) \\ &= h\big(\sqrt{\lambda}\Phi(X^*) + \sqrt{1-\lambda}\,\Psi(Y^*)\big) - h(\sqrt{\lambda}X^* + \sqrt{1-\lambda}\,Y^*). \end{aligned} \tag{8}$$

On the other hand, by Lemma 2,

$$\begin{aligned}
&\lambda\big(h(X)-h(X^*)\big)+(1-\lambda)\big(h(Y)-h(Y^*)\big)\\
&\quad=\lambda\big(h\big(\Phi(X^*)\big)-h(X^*)\big)+(1-\lambda)\big(h\big(\Psi(Y^*)\big)-h(Y^*)\big)\\
&\quad=\mathbb{E}\{\lambda\log\Phi'(X^*)+(1-\lambda)\log\Psi'(Y^*)\}. \qquad (9)
\end{aligned}$$

Thus both sides of (7) have been rewritten in terms of the Gaussian X^* and Y^*. We now compare (8) and (9). Toward this aim we make the change of variable $(X^*,Y^*)\to(\tilde{X},\tilde{Y})$ where

$$\begin{cases}\tilde{X}=\sqrt{\lambda}X^*+\sqrt{1-\lambda}\,Y^*\\ \tilde{Y}=-\sqrt{1-\lambda}\,X^*+\sqrt{\lambda}Y^*.\end{cases} \qquad (10)$$

Again $\tilde{X},\tilde{Y}$ are i.i.d. Gaussian and

$$\begin{cases}X^*=\sqrt{\lambda}\tilde{X}-\sqrt{1-\lambda}\,\tilde{Y}\\ Y^*=\sqrt{1-\lambda}\,\tilde{X}+\sqrt{\lambda}\tilde{Y}.\end{cases} \qquad (11)$$

To simplify the notation define

$$\begin{aligned}\Theta_{\tilde{y}}(\tilde{x})&=\sqrt{\lambda}\Phi(\sqrt{\lambda}\tilde{x}-\sqrt{1-\lambda}\,\tilde{y})\\&\quad+\sqrt{1-\lambda}\,\Psi(\sqrt{1-\lambda}\,\tilde{x}+\sqrt{\lambda}\tilde{y}). \qquad (12)\end{aligned}$$

Then (8) becomes

$$\begin{aligned}&h(\sqrt{\lambda}X+\sqrt{1-\lambda}\,Y)-h(\sqrt{\lambda}X^*+\sqrt{1-\lambda}\,Y^*)\\&\qquad\qquad=h\big(\Theta_{\tilde{Y}}(\tilde{X})\big)-h(\tilde{X}). \qquad (13)\end{aligned}$$

Here Lemma 2 cannot be applied directly because $\Theta_{\tilde{Y}}(\tilde{X})$ is not a deterministic function of $\tilde{X}$. But since conditioning reduces entropy,

$$h\big(\Theta_{\tilde{Y}}(\tilde{X})\big)\ge h\big(\Theta_{\tilde{Y}}(\tilde{X})\big|\tilde{Y}\big) \qquad (14)$$

Now for fixed $\tilde{y}$, since Φ and Ψ have triangular Jacobian matrices with positive diagonal elements, the Jacobian matrix of $\Theta_{\tilde{y}}$ is also triangular with positive diagonal elements. Thus, by Lemma 2,

$$h\big(\Theta_{\tilde{Y}}(\tilde{X})\big|\tilde{Y}=\tilde{y}\big)-h(\tilde{X})=\mathbb{E}\{\log\Theta'_{\tilde{y}}(\tilde{X})\} \qquad (15)$$

where $\Theta'_{\tilde{y}}$ is the Jacobian of the transformation $\Theta_{\tilde{y}}$. Since $\tilde{X}$ and $\tilde{Y}$ are independent, averaging over $\tilde{Y}$ yields

$$h\big(\Theta_{\tilde{Y}}(\tilde{X})\big|\tilde{Y}\big)-h(\tilde{X})=\mathbb{E}\{\log\Theta'_{\tilde{Y}}(\tilde{X})\}. \qquad (16)$$

Therefore, by (13)-(14)

$$\begin{aligned}&h(\sqrt{\lambda}X+\sqrt{1-\lambda}\,Y)-h(\sqrt{\lambda}X^*+\sqrt{1-\lambda}\,Y^*)\\&\qquad\qquad\ge\mathbb{E}\{\log\Theta'_{\tilde{Y}}(\tilde{X})\}. \qquad (17)\end{aligned}$$

On the other hand, (9) becomes

$$\begin{aligned}
&\lambda\big(h(X)-h(X^*)\big)+(1-\lambda)\big(h(Y)-h(Y^*)\big)\\
&\quad=\mathbb{E}\{\lambda\log\Phi'(\sqrt{\lambda}\tilde{X}-\sqrt{1-\lambda}\,\tilde{Y}) \qquad (18)\\
&\qquad+(1-\lambda)\log\Psi'(\sqrt{1-\lambda}\,\tilde{X}+\sqrt{\lambda}\tilde{Y})\}\\
&\quad=\sum_{i=1}^{n}\mathbb{E}\big\{\lambda\log\frac{\partial\Phi_i}{\partial x_i}(\sqrt{\lambda}\tilde{X}-\sqrt{1-\lambda}\,\tilde{Y}) \qquad (19)\\
&\qquad+(1-\lambda)\log\frac{\partial\Psi_i}{\partial y_i}(\sqrt{1-\lambda}\,\tilde{X}+\sqrt{\lambda}\tilde{Y})\big\}\\
&\quad\le\sum_{i=1}^{n}\mathbb{E}\log\big\{\lambda\frac{\partial\Phi_i}{\partial x_i}(\sqrt{\lambda}\tilde{X}-\sqrt{1-\lambda}\,\tilde{Y}) \qquad (20)\\
&\qquad+(1-\lambda)\frac{\partial\Psi_i}{\partial y_i}(\sqrt{1-\lambda}\,\tilde{X}+\sqrt{\lambda}\tilde{Y})\big\}\\
&\quad=\sum_{i=1}^{n}\mathbb{E}\log\frac{\partial\big(\Theta_{\tilde{Y}}\big)_i}{\partial\tilde{x}_i}(\tilde{X})=\mathbb{E}\log\Theta'_{\tilde{Y}}(\tilde{X}) \qquad (21)\\
&\quad\le h(\sqrt{\lambda}X+\sqrt{1-\lambda}\,Y)-h(\sqrt{\lambda}X^*+\sqrt{1-\lambda}\,Y^*) \qquad (22)
\end{aligned}$$

where in (20) we have used Jensen's inequality $\lambda\log a+(1-\lambda)\log b\le\log(\lambda a+(1-\lambda)b)$ on each component, in (21) the fact that the Jacobian matrix of $\Theta_{\tilde{y}}$ is triangular with positive diagonal elements, and (22) is (17). This proves (2).

IV. The Case of Equality

Equality in (2) holds if and only if both (14) and (20) are equalities. Equality in (20) holds if and only if for all $i=1,2\dots,n$,

$$\frac{\partial\Phi_i}{\partial x_i}(X^*)=\frac{\partial\Psi_i}{\partial y_i}(Y^*)\quad\text{a.e.} \qquad (23)$$

Since X^* and Y^* are independent Gaussian random vectors this implies that $\frac{\partial\Phi_i}{\partial x_i}$ and $\frac{\partial\Psi_i}{\partial y_i}$ are constant and equal. Thus in particular $\Theta'_{\tilde{y}}$ is constant. Now equality in (14) holds if and only if $\Theta_{\tilde{Y}}(\tilde{X})$ is independent of $\tilde{Y}$, thus $\Theta_{\tilde{y}}(\tilde{X})=\Theta(\tilde{X})$ does not depend on the particular value of $\tilde{y}$. Thus for all $i,j=1,2,\dots,n$,

$$\begin{aligned}0&=\frac{\partial(\Theta_{\tilde{y}}(\tilde{X}))_i}{\partial\tilde{y}_j}\\&=-\sqrt{\lambda}\sqrt{1-\lambda}\,\frac{\partial\Phi_i}{\partial x_j}(\sqrt{\lambda}\tilde{X}-\sqrt{1-\lambda}\,\tilde{Y})\\&\quad+\sqrt{1-\lambda}\sqrt{\lambda}\frac{\partial\Psi_i}{\partial y_j}(\sqrt{1-\lambda}\,\tilde{X}+\sqrt{\lambda}\tilde{Y}) \qquad (24)\end{aligned}$$

which implies

$$\frac{\partial\Phi_i}{\partial x_j}(X^*)=\frac{\partial\Psi_i}{\partial y_j}(Y^*)\quad\text{a.e.}, \qquad (25)$$

hence $\frac{\partial\Phi_i}{\partial x_j}$ and $\frac{\partial\Psi_i}{\partial y_j}$ are constant and equal for any $i,j=1,2,\dots,n$. Therefore, Φ and Ψ are linear transformations, equal up to an additive constant. It follows that $\Phi(X^*)$ and $\Phi(Y^*)$ (hence X and Y) are Gaussian with identical covariances. This ends the proof of the Theorem. □

Extensions of similar ideas when X^*,Y^* need not be Gaussian can be found in [32].

Appendix A

The differential entropy $h(X)=-\int f\log f$ of a random vector X with density f is not always well-defined because the negative and positive parts of the integral might be both infinite, as in the example $f(x)=1/(2x\log^2 x)$ for $0<x<1/\mathrm{e}$ and $\mathrm{e}<x<+\infty$, and $=0$ otherwise [12].

Proposition 1: Let X be an random vector with density f and finite second moments. Then $h(X)=-\int f\log f$ is well-defined and $<+\infty$.

Proof: Let Z be any Gaussian vector with density $g>0$. On one hand, since X has finite second moments, the integral

$\int f \log g$ is finite. On the other hand, since g never vanishes, the probability measure of X is absolutely continuous with respect to that of Z. Therefore, the divergence $D(f\|g)$ is equal to the integral $\int f \log(f/g)$. Since the divergence is non-negative, it follows that $-\int f \log f = -\int f \log g - D(f\|g) \leq -\int f \log g$ is well-defined and $< +\infty$ (the positive part of the integral is finite). $\square$

APPENDIX B

It is stated in [33, Appendix II A] that strong smoothness properties of distributions of $Y = X + Z$ for independent Gaussian Z are "very well known in certain mathematical circles" but it seems difficult to find a reference.

The following result is stated for an arbitrary random vector X. It is not required that X have a density. It could instead follow e.g., a discrete distribution.

Proposition 2: Let X be any random vector and Z be any independent Gaussian vector with density $g > 0$. Then $Y = X + Z$ has a bounded, positive, indefinitely differentiable (hence continuous) density that tends to zero at infinity, whose all derivatives are also bounded and tend to zero at infinity.

Proof: Taking characteristic functions, $\mathbb{E}(e^{it\cdot Y}) = \mathbb{E}(e^{it\cdot X}) \cdot \mathbb{E}(e^{it\cdot Z})$, where $\hat{g}(t) = \mathbb{E}(e^{it\cdot Z})$ is the Fourier transform of the Gaussian density g. Now $\hat{g}(t)$ is also a Gaussian function with exponential decay at infinity and $|\mathbb{E}(e^{it\cdot Y})| \leq |\mathbb{E}(e^{it\cdot X})|\cdot|\mathbb{E}(e^{it\cdot Z})| \leq \mathbb{E}(|e^{it\cdot X}|)\cdot|\hat{g}(t)| = |\hat{g}(t)|$. Therefore, the Fourier transform of the probability measure of Y (which is always continuous) also has exponential decay at infinity. In particular, this Fourier transform is integrable, and by the Riemann-Lebesgue lemma, Y has a bounded continuous density which tends to zero at infinity. Similarly, for any monomial[5] t^α, $(it)^\alpha \mathbb{E}(e^{it\cdot Y})$ is integrable and is the Fourier transform of the αth partial derivative of the density of Y, which is, therefore, also bounded continuous and tends to zero at infinity.

It remains to prove that the density of Y is positive. Let Z_1, Z_2 be independent Gaussian random vectors with density ϕ equal to that of $Z/\sqrt{2}$ so that Z has the same distribution as $Z_1 + Z_2$. By what has just been proved, $X + Z_1$ follows a continuous density f. Since Y has the same distribution as $(X + Z_1) + Z_2$, its density is equal to the convolution product $f * \phi\,(y) = \int_{\mathbb{R}} \phi(z) f(y-z)\,\mathrm{d}z$. Now ϕ is positive, and for any $y \in \mathbb{R}^n$, $\int_{\mathbb{R}} \phi(z) f(y-z)\,\mathrm{d}z = 0$ would imply that f vanishes identically, which is impossible. $\square$

APPENDIX C
FIRST PROOF OF LEMMA 1

We use the notation f for densities (p.d.f.'s). In the first dimension, for each $x_1^* \in \mathbb{R}$, define $\Phi_1(x_1^*)$ such that

$$\int_{-\infty}^{\Phi_1(x_1^*)} f_{X_1} = \int_{-\infty}^{x_1^*} f_{X_1^*}. \tag{26}$$

Since the densities are continuous and positive, Φ_1 is continuously differentiable and increasing; differentiating gives

$$f_{X_1}(\Phi_1(x_1^*))\,\frac{\partial \Phi_1}{\partial x_1^*}(x_1^*) = f_{X_1^*}(x_1^*) \tag{27}$$

which proves the result in one dimension: X_1 has the same distribution as $\Phi_1(X_1^*)$ where $\dfrac{\partial \Phi_1}{\partial x_1^*}$ is positive.

In the first two dimensions, for each x_1^*, x_2^* in $\mathbb{R}$, define $\Phi_2(x_1^*, x_2^*)$ such that

$$\int_{-\infty}^{\Phi_2(x_1^*,x_2^*)} f_{X_1,X_2}(\Phi_1(x_1^*),\,\cdot\,)\,\frac{\partial \Phi_1}{\partial x_1^*}(x_1^*) = \int_{-\infty}^{x_2^*} f_{X_1^*,X_2^*}(x_1^*,\,\cdot\,). \tag{28}$$

Again Φ_2 is continuously differentiable and increasing in x_2^*; differentiating gives

$$f_{X_1,X_2}(\Phi_1(x_1^*), \Phi_2(x_1^*, x_2^*))\,\frac{\partial \Phi_1}{\partial x_1^*}(x_1^*)\frac{\partial \Phi_2}{\partial x_2^*}(x_1^*, x_2^*) = f_{X_1^*,X_2^*}(x_1^*, x_2^*) \tag{29}$$

which proves the result in two dimensions. Continuing in this manner we arrive at

$$\begin{aligned} &f_{X_1,X_2,\ldots,X_n}(\Phi_1(x_1^*), \Phi_2(x_1^*, x_2^*), \ldots, \Phi_n(x_1^*, x_2^*, \ldots, x_n^*)) \\ &\quad\times \frac{\partial \Phi_1}{\partial x_1^*}(x_1^*)\frac{\partial \Phi_2}{\partial x_2^*}(x_1^*, x_2^*)\cdots\frac{\partial \Phi_n}{\partial x_n^*}(x_1^*, x_2^*, \ldots, x_n^*) \\ &= f_{X_1^*,X_2^*,\ldots,X_n^*}(x_1^*, x_2^*, \ldots, x_n^*) \end{aligned} \tag{30}$$

which shows that $X = (X_1, X_2, \ldots, X_n)$ has the same distribution as $\Phi(X_1^*, X_2^*, \ldots, X_n^*) = \big(\Phi_1(X_1^*), \Phi_2(X_1^*, X_2^*), \ldots, \Phi_n(X_1^*, X_2^*, \ldots, X_n^*)\big)$. The Jacobian matrix of Φ has the form

$$\mathbf{J}_\Phi(x_1^*, x_2^*, \ldots, x_n^*) = \begin{pmatrix} \frac{\partial \Phi_1}{\partial x_1^*} & 0 & \cdots & 0 \\ \frac{\partial \Phi_2}{\partial x_1^*} & \frac{\partial \Phi_2}{\partial x_2^*} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial \Phi_n}{\partial x_1^*} & \frac{\partial \Phi_n}{\partial x_2^*} & \cdots & \frac{\partial \Phi_n}{\partial x_n^*} \end{pmatrix} \tag{31}$$

where all diagonal elements are positive since by construction each Φ_k is increasing in x_k^*. $\square$

APPENDIX D
SECOND PROOF OF LEMMA 1

We use the notation F for distribution functions (c.d.f.'s). We also note $F_{X_2|X_1}(x_2|x_1) = \mathbb{P}(X_2 \leq x_2 \mid X_1 = x_1)$ and let $F^{-1}_{X_2|X_1}(\cdot|x_1)$ be the corresponding inverse function in the argument x_2 for a fixed value of x_1. Such inverse functions are well-defined since it is assumed that X is a random vector with continuous, positive density.

The inverse transform sampling method is well known for univariate random variables but its multivariate generalization is not.

Lemma 3 (Multivariate Inverse Transform Sampling Method (see, e.g., [34, Algorithm 2])): Let $U = (U_1, U_2, \ldots, U_n)$ be uniformly distributed on $[0, 1]^n$. The vector $\Phi(U)$ with components

$$\begin{aligned} \Phi_1(U_1) &= F^{-1}_{X_1}(U_1) \\ \Phi_2(U_1, U_2) &= F^{-1}_{X_2|X_1}(U_2|\Phi_1(U_1)) \\ &\;\;\vdots \\ \Phi_n(U_1, U_2, \ldots, U_n) &= \\ F^{-1}_{X_n|X_1,\ldots,X_{n-1}}&(U_n|\Phi_1(U_1), \ldots, \Phi_{n-1}(U_1, \ldots, U_{n-1})) \end{aligned} \tag{32}$$

has the same distribution as X.

[5] Here we use the multi-index notation $t^\alpha = t_1^{\alpha_1} t_1^{\alpha_1} \cdots t_n^{\alpha_n}$.

Proof: By inverting Φ, it is easily seen that an equivalent statement is that the random vector $\big(F_{X_1}(X_1), F_{X_2|X_1}(X_2|X_1), \ldots, F_{X_n|X_1,\ldots,X_{n-1}}(X_n|X_1,\ldots,X_{n-1})\big)$ is uniformly distributed in $[0,1]^n$. Clearly $F_{X_1}(X_1)$ is uniformly distributed in $[0,1]$, since

$$\begin{aligned}\mathbb{P}(F_{X_1}(X_1) \leq u_1) &= \mathbb{P}(X_1 \leq F_{X_1}^{-1}(u_1)) \\ &= F_{X_1} \circ F_{X_1}^{-1}(u_1) \\ &= u_1.\end{aligned} \tag{33}$$

Similarly, for any $k > 0$ and fixed $x_1, \ldots, x_{k-1}$, $F_{X_k|X_1,\ldots,X_{k-1}}(X_k|X_1 = x_1, X_2 = x_2, \ldots, X_{k-1} = x_{k-1})$ is also uniformly distributed in $[0,1]$. The result follows by the chain rule. $\square$

Proof of Lemma 1: By Lemma 3, X has the same distribution as $\Phi(U)$, where each $\Phi_k(u_1, u_2, \ldots, u_k)$ is increasing in u_k. Similarly X^* has the same distribution as $\Psi(U)$, where both Φ and Ψ have (lower) triangular Jacobian matrices $\mathbf{J}_\Phi, \mathbf{J}_\Psi$ with positive diagonal elements. Then X has the same distribution as $\Phi(\Psi^{-1}(X^*))$. By the chain rule for differentiation, the transformation $\Phi \circ \Psi^{-1}$ has Jacobian matrix $(\mathbf{J}_\Phi \circ \Psi^{-1}) \cdot \mathbf{J}_{\Psi^{-1}} = (\mathbf{J}_\Phi \circ \Psi^{-1}) \cdot (\mathbf{J}_\Psi \circ \Psi^{-1})^{-1}$. This product of (lower) triangular matrices with positive diagonal elements and is again (lower) triangular with positive diagonal elements. $\square$

ACKNOWLEDGMENT

The author would like to thank Max Costa, Tom Courtade and Cédric Villani for their discussions, Tamás Linder for pointing out references [16]–[18] in connection with (3) and the anonymous reviewers for pointing out references [28]–[30] related to optimal transport theory.

REFERENCES

[1] C. E. Shannon, "A mathematical theory of communication," *Bell System Tech. J.*, vol. 27, no. 4, pp. 623–656, Oct. 1948.

[2] A. J. Stam, "Some inequalities satisfied by the quantities of information of Fisher and Shannon," *Inf. Control*, vol. 2, no. 2, pp. 101–112, Jun. 1959.

[3] N. M. Blachman, "The convolution inequality for entropy powers," *IEEE Trans. Inf. Theory*, vol. 11, no. 2, pp. 267–271, Apr. 1965.

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

[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] E. A. Carlen and A. Soffer, "Entropy production by block variable summation and central limit theorems," *Commun. Math. Phys.*, vol. 140, no. 2, pp. 339–371, 1991.

[7] S. J. Szarek and D. Voiculescu, "Shannon's entropy power inequality via restricted Minkowski sums," in *Geometric Aspects of Functional Analysis* (Lecture Notes in Mathematics), vol. 1745. Berlin, Germany: Springer, 2000, pp. 257–262.

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

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

[10] J. Binia, "On divergence-power inequalities," *IEEE Trans. Inf. Theory*, vol. 53, no. 3, pp. 1179–1182, Mar. 2007.

[11] O. Rioul, "A simple proof of the entropy-power inequality via properties of mutual information," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Nice, France, Jun. 2007, pp. 46–50.

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

[13] L. Wang and M. Madiman, "Beyond the entropy power inequality, via rearrangements," *IEEE Trans. Inf. Theory*, vol. 60, no. 9, pp. 5116–5137, Sep. 2014.

[14] M. Madiman, J. Melbourne, and P. Xu. (Apr. 2016). "Forward and reverse entropy power inequalities in convex geometry." [Online]. Available: https://arxiv.org/abs/1604.04225

[15] S. G. Bobkov and G. P. Chistyakov, "Entropy power inequality for the Rényi entropy," *IEEE Trans. Inf. Theory*, vol. 61, no. 2, pp. 708–714, Feb. 2015.

[16] J. Binia, M. Zakai, and J. Ziv, "On the ϵ-entropy and the rate-distortion function of certain non-Gaussian processes," *IEEE Trans. Inf. Theory*, vol. 20, no. 4, pp. 517–524, Jul. 1974.

[17] A. R. Barron, "Entropy and the central limit theorem," *Ann. Probab.*, vol. 14, no. 1, pp. 336–342, 1986.

[18] T. Linder and R. Zamir, "On the asymptotic tightness of the Shannon lower bound," *IEEE Trans. Inf. Theory*, vol. 40, no. 6, pp. 2026–2031, Nov. 1994.

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

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

[21] O. Rioul and M. H. M. Costa, "On some almost properties," in *Proc. IEEE Inf. Theory Appl. Workshop (ITA)*, San Diego, USA, Feb. 2016, pp. 1–5.

[22] H. Knöthe, "Contributions to the theory of convex bodies," *Michigan Math. J.*, vol. 4, no. 1, pp. 39–52, 1957.

[23] V. D. Milman and G. Schechtman, "Asymptotic theory of finite dimensional normed spaces," *Lecture Notes in Mathematics* (Lecture Notes in Mathematics), vol. 1200. Berlin, Germany: Springer, 1986.

[24] A. A. Giannopoulos and V. D. Milman, "Asymptotic convex geometry: Short overview," in *Different Faces of Geometry*, S. Donaldson, Y. Eliashberg, and M. Gromov, Eds., vol. 3. Berlin, Germany: Springer, 2004, pp. 87–162.

[25] S. Artstein-Avidan, A. Giannopoulos, and V. D. Milman, "Asymptotic geometric analysis I," *Mathematical Surveys and Monographs*, vol. 202, Providence, RI, USA: AMS, 2015.

[26] C. Villani, *Topics in Optimal Transportation*. Providence, RI, USA: AMS, 2003.

[27] C. Villani, "Optimal transport: Old and new," in *Science and Business Media*, vol. 338. Berlin, Germany: Springer, 2008.

[28] F. Barthe, "Optimal Young's inequality and its converse: A simple proof," *Geometric Funct. Anal. GAFA*, vol. 8, no. 2, pp. 234–242, 1998.

[29] D. Cordero-Erausquin, B. Nazaret, and C. Villani, "A mass-transportation approach to sharp Sobolev and Gagliardo–Nirenberg inequalities," *Adv. Math.*, vol. 182, no. 2, pp. 307–332, 2004.

[30] D. Cordero-Erausquin, R. J. McCann, and M. Schmuckenschläger, "Prekopa-Leindler type inequalities on Riemannian manifolds, Jacobi fields, and optimal transport," *Ann. Faculté Sci. Toulouse, Math.*, vol. 15, no. 4, pp. 613–635, 2006.

[31] T. A. Courtade, M. Fathi, and A. Pananjady. (Oct. 2016). "Wasserstein stability of the entropy power inequality for log-concave densities." [Online]. Available: https://arxiv.org/abs/1610.07969

[32] O. Rioul, "Optimal transportation to the entropy-power inequality," in *Proc. IEEE Inf. Theory Appl. Workshop (ITA)*, San Diego, CA, USA, Feb. 2017.

[33] 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.

[34] O. Caster and L. Ekenberg "Combining second-order belief distributions with qualitative statements in decision analysis," in *Managing Safety of Heterogeneous Systems* (Lecture Notes in Economics and Mathematical Systems), vol. 658, Y. Ermoliev, M. Makowski, and K. Marti, Eds. Berlin, Germany: Springer, 2012.

Olivier Rioul is full Professor at the Department of Communication and Electronics at Telecom ParisTech, Université Paris-Saclay, France, and also Professor chargé de cours at the Department of Applied Mathematics at École Polytechnique, Université Paris-Saclay, France. He graduated from École Polytechnique, Paris, France in 1987 and from École Nationale Supérieure des Télécommunications, Paris, France in 1989. He obtained his PhD degree from École Nationale Supérieure des Télécommunications, Paris, France in 1993. He is also the father of five beautiful children.