

Article

# Onicescu's Informational Energy and Correlation Coefficient in Exponential Families

Frank Nielsen 

Sony Computer Science Laboratories Inc., Tokyo 141-0022, Japan; Frank.Nielsen@acm.org

**Abstract:** The informational energy of Onicescu is a positive quantity that measures the amount of uncertainty of a random variable. However, contrary to Shannon's entropy, the informational energy is strictly convex and increases when randomness decreases. We report a closed-form formula for Onicescu's informational energy and its associated correlation coefficient when the probability distributions belong to an exponential family. We show how to instantiate the generic formula for several common exponential families. Finally, we discuss the characterization of valid thermodynamic process trajectories on a statistical manifold by enforcing that the entropy and the informational energy shall vary in opposite directions.

**Keywords:** Onicescu's informational energy; Shannon's entropy; Onicescu's correlation coefficient; Rényi's quadratic entropy; cauchy-schwarz divergence; Simpson's diversity index; exponential families; location-scale families; thermodynamics process

## 1. Introduction

### 1.1. Onicescu's Informational Energy

Let  $(\mathcal{X}, \mathcal{F}, \mu)$  be a probability space [1] with  $\sigma$ -algebra  $\mathcal{F}$  on the sample space  $\mathcal{X}$ , and  $\mu$  a base measure often chosen as the Lebesgue measure or as the counting measure. Let  $\mathcal{M}$  denote the set of Radon-Nikodym densities of probability measures dominated by  $\mu$ . Two probability densities  $p$  and  $q$  are said equal (i.e.,  $p = q$ ) if and only if  $p(x) = q(x)$   $\mu$ -almost everywhere, and different (i.e.,  $p \neq q$ ) when  $\mu(\{x \in \mathcal{X} : p(x) \neq q(x)\}) \neq 0$ .

Octav Onicescu [2,3] (1892–1983) was a renowned Romanian mathematician who founded the school of probability theory and statistics [4] in Romania. Onicescu introduced the informational energy [5] (also termed information energy [6] in the literature) of a probability measure  $P \ll \mu$  with Radon-Nikodym density  $p = \frac{dP}{d\mu}$  as

$$I(p) := \int p^2(x) d\mu(x) > 0. \quad (1)$$

The Rényi entropy [7,8] of order-2 can be written using the informational energy:

$$R_2(p) := -\log \int p^2(x) d\mu(x) = -\log I(p), \quad (2)$$

as well as Vajda's quadratic entropy [9]:

$$V_2(p) := 1 - \int p^2(x) d\mu(x) = 1 - I(p). \quad (3)$$

Notice that it follows from Cauchy-Schwarz's inequality that we have  $I(p) \leq 1$  when  $p \in L^2(\mu)$ , the Lebesgue space of square integrable functions [1]. The informational energy for continuous distributions  $p$  is  $I(p) = \int_{\mathcal{X}} p^2(x) dx$  and the informational energy for discrete distributions  $q$ ,  $I(q) = \sum_{x \in \mathcal{X}} q^2(x)$ . Notice that the informational energy in the continuous case is not a limit of the informational energy in the discrete case [6,10].



**Citation:** Nielsen, F. Onicescu's Informational Energy and Correlation Coefficient in Exponential Families. *Foundations* **2022**, *2*, 362–376. <https://doi.org/10.3390/foundations2020025>

Academic Editor: Donald J. Jacobs

Received: 4 February 2022

Accepted: 16 April 2022

Published: 19 April 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

The informational energy is an important concept in statistics which fruitfully interplays with Shannon’s entropy [11,12]  $H(p)$ :

$$H(p) := \int p(x) \log \frac{1}{p(x)} d\mu(x) = \int -p(x) \log p(x) d\mu(x). \tag{4}$$

Notice that the informational energy is always positive but the Shannon’s entropy may be negative for continuous distributions (e.g., differential entropy of the normal distributions for small standard deviation). For the Dirac’s distribution  $\delta_e$  (with  $\delta_e(x) = 1$  when  $x = e$  and 0 otherwise), the informational energy is  $I(\delta_e) = +\infty$  but Shannon’s entropy is  $H(\delta_e) = -\infty$ .

The informational energy measures the amount of uncertainty of a random variable like Shannon’s entropy but augments when randomness decreases.

Onicescu’s informational energy finds also applications in various other fields since Onicescu’s informational energy corresponds to the Herfindahl–Hirschman index [13]  $\text{HHI}(p) = I(p)$  in economics, to the index coincidence [14]  $\text{IC}(p) = I(p)$  in information theory (originally developed for cryptanalysis [15]), and is related to Simpson’s diversity index [16,17]  $S(p) = I(p)$  in ecology (also called the Gini-Simpson index of diversity [18]). Section 4 will consider the joint variations of the Shannon entropy and the informational energy to characterize valid thermodynamic process paths on a statistical manifold.

Another key difference with Shannon’s entropy is that Shannon’s entropy is always strictly concave [12] but the informational energy is always strictly convex:

**Property 1.** *Onicescu’s informational energy  $I(\cdot)$  is a strictly convex functional.*

**Proof.** A function  $F$  is strictly convex if and only if for any  $\alpha \in (0,1)$  and  $p \neq q$  two densities of a convex domain  $\mathcal{M}$ , we have  $F((1 - \alpha)p + \alpha q) < (1 - \alpha)F(p) + \alpha F(q)$ . Let us check this strict inequality for the informational energy  $I$ :

$$I((1 - \alpha)p + \alpha q) = (1 - \alpha)^2 I(p) + \alpha^2 I(q) + 2\alpha(1 - \alpha) \int p(x)q(x) d\mu(x), \tag{5}$$

$$= (1 - \alpha)I(p) + \alpha I(q) + 2\alpha(1 - \alpha) \int p(x)q(x) d\mu(x) - \alpha(1 - \alpha)(I(p) + I(q)),$$

$$= (1 - \alpha)I(p) + \alpha I(q) - \underbrace{\alpha(1 - \alpha) \int (p(x) - q(x))^2 d\mu(x)}_{>0}, \tag{6}$$

$$< (1 - \alpha)I(p) + \alpha I(q), \tag{7}$$

when  $p \neq q$  and  $\alpha \in (0,1)$ .  $\square$

Table 1 summarizes the comparison between Shannon’s entropy and Onicescu’s informational energy.

Since  $I(p)$  is strictly convex, we can define the informational energy divergence as the following Jensen divergence [19] measuring the convexity gap:

$$J_I(p, q) := \frac{I(p) + I(q)}{2} - I\left(\frac{p + q}{2}\right), \tag{8}$$

$$= \frac{1}{4} \int_{\mathcal{X}} (p(x) - q(x))^2 d\mu(x). \tag{9}$$

For a uniform discrete distribution  $p$  on an alphabet  $\mathcal{X}$  of  $d$  letters, we have  $I(p) = \frac{1}{d}$ , and for any probability mass function  $p$  on  $\mathcal{X}$ , we have  $I(p) \geq \frac{1}{d}$ . Recall that  $H(p) \leq \log d$  with equality when  $p$  is the discrete uniform distribution. More generally, for a continuous density on an interval  $\mathcal{X} = [a, b]$ , we have  $I(p) \geq \frac{1}{b-a}$ .

For discrete or continuous distributions, we have the following inequality [20] (Proposition 5.8.5):

$$H(p) + \frac{1}{2}I(p) \geq 1 - \log 2 > 0.69897, \tag{10}$$

and Shannon’s cross-entropy

$$H^\times(p : q) = \int -p(x) \log q(x) d\mu(x) \tag{11}$$

can be lower bounded using the informational energy as follows (Problem 5.8 in [20]): For any  $x > 0$ , we have  $\log x \leq x - 1$ . Thus we get  $\log q(x) \leq q(x) - 1$  and  $-p(x) \log q(x) \geq p(x) - p(x)q(x)$ . Therefore we have

$$H^\times(p : q) \geq 1 - \int p(x)q(x) d\mu(x).$$

Using Cauchy–Schwarz’s inequality

$$\int p(x)q(x) d\mu(x) \leq \sqrt{\int p(x)^2 d\mu(x)} \sqrt{\int q(x)^2 d\mu(x)},$$

we get:

$$H^\times(p : q) \geq 1 - \sqrt{I(p) I(q)}. \tag{12}$$

In particular, we get a lower bound on Shannon’s entropy:

$$H(p) = H^\times(p : q) \geq 1 - I(p).$$

**Table 1.** Comparison between Shannon’s entropy and Onicescu’s informational energy.

	Entropy $H(p)$	Informational Energy $I(p)$
convexity	$H(p) = - \int p(x) \log p(x) d\mu(x)$	$I(p) = \int p^2(x) d\mu(x)$
range	strictly concave	strictly convex
uncertainty measure	can be negative	always positive
uniform discrete distribution $u$	augments with disorder	decreases with disorder
(with alphabet size $ \mathcal{X}  = d$ )	$H(u) = \log d$	$I(u) = \frac{1}{d}$
bound	$H(p) \geq 1 - I(p)$	$I(p) \geq 1 - H(p)$
	Inequality: $H(p) + \frac{1}{2} I(p) \geq 1 - \log 2$	

For an in-depth treatment of Onicescu’s informational energy, we refer to the paper [6] (77 pages, with main properties listed in pp. 167–169 as statistical applications). Onicescu’s informational energy has been used in physics [21], information theory in electronic structure theory of atomic and molecular systems [22–24], machine learning [17,25], and complex systems [26], among others.

1.2. Onicescu’s Correlation Coefficient

Onicescu also defined a correlation coefficient (see [20], Chapter 5):

$$\rho(p, q) := \frac{I(p, q)}{\sqrt{I(p) I(q)}}, \tag{13}$$

where  $I(p, q)$  denotes the cross-informational energy:

$$I(p, q) := \int p(x)q(x) d\mu(x), \tag{14}$$

with  $I(p) = I(p, p)$ . Notice that it follows from the Cauchy-Schwarz inequality that  $I(p, q) \leq \sqrt{I(p)I(q)}$ , and therefore we have:

$$0 < \rho(p, q) \leq 1,$$

assuming both densities  $p$  and  $q$  belong to the Lebesgue space  $L^2(\mu)$ .

Notice that the informational energy of a statistical mixture  $m(x) = \sum_{i=1}^k w_i p_i(x)$  with  $k$  weighted components  $p_1(x), \dots, p_k(x)$  (with  $w \in \Delta_k$  the  $(k - 1)$ -dimensional standard simplex) can be expressed as follows:

$$I(m) = \int \left( \sum_{i=1}^k w_i p_i(x) \right)^2 d\mu(x) = \sum_{i=1}^k \sum_{j=1}^k w_i w_j I(p_i, p_j). \tag{15}$$

The Cauchy-Schwarz divergence [27,28] is defined by

$$D_{CS}(p, q) := -\log \left( \frac{\int_{\mathcal{X}} p(x)q(x)d\mu(x)}{\sqrt{(\int_{\mathcal{X}} p(x)^2 d\mu(x))(\int_{\mathcal{X}} q(x)^2 d\mu(x))}} \right) \geq 0. \tag{16}$$

Thus, the Cauchy-Schwarz divergence is a projective divergence (that is, we have  $D_{CS}(p, q) = D_{CS}(\lambda p, \lambda' q)$  for any  $\lambda > 0$  and  $\lambda' > 0$ ) which can be rewritten using the Onicescu’s correlation coefficient as:

$$D_{CS}(p, q) = -\log(\rho(p, q)). \tag{17}$$

### 1.3. Exponential Families

Consider a natural exponential family [29,30] (NEF)

$$\mathcal{E} = \left\{ p_{\theta}(x) = \exp\left(\theta^{\top} t(x) - F(\theta) + k(x)\right) : \theta \in \Theta \right\}, \tag{18}$$

where  $t(x)$  denotes the minimal sufficient statistics,  $k(x)$  an auxiliary measure carrier term, and

$$F(\theta) := \log \left( \int_{\mathcal{X}} \exp(\theta^{\top} t(x)) d\mu(x) \right), \tag{19}$$

the cumulant function which is commonly called the log-normalizer (or log-partition function in statistical physics). Parameter  $\theta$  is called the natural parameter and is defined on the open convex natural parameter space  $\Theta$ .

Many familiar families of distributions  $\{p_{\lambda}(x) \mid \lambda \in \Lambda\}$  are exponential families in disguise after reparameterization:  $p_{\lambda}(x) = p_{\theta(\lambda)}(x)$  (e.g., normal family or Poisson family). Those families are called exponential families (omitting the leading adjective ‘natural’), and their densities are canonically factorized as follows:

$$p_{\lambda}(x) = \exp\left(\theta(\lambda)^{\top} t(x) - F(\theta(\lambda)) + k(x)\right). \tag{20}$$

We call parameter  $\lambda \in \Lambda$  the source parameter, and parameter  $\theta(\lambda) \in \Theta$  is called the corresponding natural parameter. Densities of an exponential family have all the same support  $\mathcal{X}$ .

## 2. Onicescu’s Informational Energy and Correlation Coefficient in Exponential Families

In this section, we first report a closed-form formula for the informational energy and its associated correlation coefficient in Section 2.1. We then describe some statistical divergences related to Onicescu’s correlation coefficient in Section 2.2.

2.1. Closed-Form Formula

We report closed-form formulas for Onicescu’s informational energy and correlation coefficient when densities belong to a prescribed exponential family, and then illustrate those formula on common families of probability distributions.

**Theorem 1** (Onicescu’s informational energy and correlation coefficient in exponential families). *In an exponential family  $\mathcal{E} = \{p_\theta\}_{\theta \in \Theta}$ , Onicescu’s informational energy of a probability density  $p_\theta$  is given by:*

$$I(p_\theta) = \exp(F(2\theta) - 2F(\theta)) E_{p_{2\theta}}[\exp(k(x))], \tag{21}$$

provided that  $2\theta \in \Theta$  so that  $p_{2\theta} \in \mathcal{E}$ . When the auxiliary carrier term  $k(x)$  vanishes, we have  $E_{p_{2\theta}}[\exp(k(x))] = 1$ .

The Onicescu’s correlation coefficient  $\rho(p_{\theta_1}, p_{\theta_2})$  between densities  $p_1 = p_{\theta_1}$  and  $p_2 = p_{\theta_2}$  is

$$\rho(p_{\theta_1}, p_{\theta_2}) = \exp(-J_F(2\theta_1 : 2\theta_2)) \times \frac{E_{p_{\theta_1+\theta_2}}[\exp(k(x))]}{\sqrt{E_{p_{2\theta_1}}[\exp(k(x))] E_{p_{2\theta_2}}[\exp(k(x))]}}, \tag{22}$$

provided that  $\theta_1 + \theta_2 \in \Theta, 2\theta_1 \in \Theta, 2\theta_2 \in \Theta$ , where

$$J_F(\theta_1, \theta_2) := \frac{F(\theta_1) + F(\theta_2)}{2} - F\left(\frac{\theta_1 + \theta_2}{2}\right) \geq 0, \tag{23}$$

is a Jensen divergence [19] induced by the cumulant function of the exponential family.

**Proof.** The proof follows the same line of arguments as in [31]. Consider the term  $I(p_{\theta_1}, p_{\theta_2})$ :

$$\begin{aligned} I(p_{\theta_1}, p_{\theta_2}) &:= \int \exp(t(x)^\top \theta_1 - F(\theta_1) + k(x)) \exp(t(x)^\top \theta_2 - F(\theta_2) + k(x)) d\mu(x) \tag{24} \\ &= \int \exp\left(t(x)^\top (\theta_1 + \theta_2) - F(\theta_1 + \theta_2) + k(x) \right. \\ &\quad \left. + F(\theta_1 + \theta_2) - F(\theta_1) - F(\theta_2) + k(x)\right) d\mu(x), \\ &= \exp(F(\theta_1 + \theta_2) - F(\theta_1) - F(\theta_2)) \int p_{\theta_1+\theta_2}(x) \exp(k(x)) d\mu(x), \tag{25} \\ &= \exp(F(\theta_1 + \theta_2) - F(\theta_1) - F(\theta_2)) E_{p_{\theta_1+\theta_2}}[\exp(k(x))], \tag{26} \end{aligned}$$

provided that  $\theta_1 + \theta_2 \in \Theta$ . This condition is always satisfied when the natural parameter space is either a cone [31] (e.g., Gaussian family, Wishart family, etc) or an affine space [32] (e.g., Poisson family, isotropic Gaussian family, etc). Since  $I(p) = I(p, p)$  and  $\rho(p, q) = \frac{I(p,q)}{\sqrt{I(p)I(q)}}$ , we deduce formula of Equations (21) and (22). □

2.2. Divergences Related to Onicescu’s Correlation Coefficient

Since  $D_{CS}(p, q) = -\log(\rho(p, q))$ , we get the following closed-form for the Cauchy-Schwarz divergence:

$$D_{CS}(p_{\theta_1}, p_{\theta_2}) = J_F(2\theta_1 : 2\theta_2) + \log\left(\frac{\sqrt{E_{p_{2\theta_1}}[\exp(k(x))] E_{p_{2\theta_2}}[\exp(k(x))]}{E_{p_{\theta_1+\theta_2}}[\exp(k(x))]} \right). \tag{27}$$

We check that when  $\theta_1 = \theta_2$ , we have  $D_{CS}(p_{\theta_1}, p_{\theta_2}) = 0$ . Closed-form formula were also reported for the Cauchy-Schwarz divergence between densities of an exponential family in [33]. Table 2 reports the formula for Onicescu’s informational energy and the Shannon’s entropy [34] for densities belonging to some common exponential families. These formula can be recovered easily from the generic formula using the canonical decom-

positions of exponential families reported in [29]. Shannon’s entropy [34] of a density  $p_\theta$  of  $\mathcal{E}$  is

$$H(p_\theta) = F(\theta) - \theta^\top \nabla F(\theta) - E_{p_\theta}[k(x)], \tag{28}$$

$$H(p_\theta) = -F^*(\eta) - E_{p_\theta}[k(x)], \tag{29}$$

where  $F^*$  denotes the Legendre-Fenchel convex conjugate and  $\eta = \nabla F(\theta) = E_{p_\theta}[t(x)]$  the moment parameter.

**Table 2.** Comparisons between Shannon’s entropy and Onicescu’s informational energy for common distributions of exponential families.  $|\cdot|$  denotes the matrix determinant,  $\Gamma(\cdot)$  the gamma function,  $\psi(\cdot)$  the digamma function, and  $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ .

Family	Entropy	Informational Energy
Generic $\mathcal{E}$	$F(\theta) - \theta^\top \nabla F(\theta) - E_{p_\theta}[k(x)]$	$e^{F(2\theta) - 2F(\theta)} E_{p_{2\theta}}[\exp(k(x))]$
Univar. normal $N(\mu, \sigma)$	$\frac{1}{2} \log(2\pi e \sigma^2)$	$\frac{1}{2\sigma\sqrt{\pi}}$
Multivar. normal $N(\mu, \Sigma)$	$\frac{1}{2} \log  2\pi e \Sigma $	$\pi^{-\frac{d}{2}} 2^{-d}  \Sigma ^{-\frac{1}{2}}$
LogNormal( $\mu, \sigma$ )	$\log(\sigma e^{\mu + \frac{1}{2}\sqrt{2\pi}})$	$\frac{1}{2\sigma\sqrt{\pi}} \exp(\frac{\sigma^2}{4} - \mu)$
Exponential( $\lambda$ )	$1 - \log \lambda$	$\frac{\lambda}{2}$
Pareto $_k(a)$	$1 + \frac{1}{a} + \log \frac{k}{a}$	$\frac{a^2}{k(2a+1)}$
Gamma( $\alpha, \beta$ )	$\alpha + \log \frac{\Gamma(\alpha)}{\beta} + (1 - \alpha)\psi(\alpha)$	$\frac{1}{\beta(2\alpha-1)B(\alpha, \frac{1}{2})}$
Beta( $\alpha, \beta$ )	$\log B(\alpha, \beta) - (\alpha - 1)\psi(\alpha) - (\beta - 1)\psi(\beta) + (\alpha + \beta - 2)\psi(\alpha + \beta)$	$B^2(\alpha, \beta) \frac{\Gamma(2\alpha-1)\Gamma(2\beta-1)}{\Gamma(2\alpha+2\beta-2)}$
Poisson( $\lambda$ )	$\lambda(1 - \log \lambda) + e^{-\lambda} \sum_{i=0}^{\infty} \frac{\lambda^i \log i!}{i!}$	$\exp(-2\lambda) \sum_{i=0}^{\infty} \frac{\lambda^{2i}}{(i!)^2}$

Notice that when  $k(x) = 0$  (no auxiliary carrier term, e.g., Gaussian family), we have  $E_p[e^{k(x)}] = E_p[1] = \int p(x) d\mu(x) = 1$  for any density  $p \in \mathcal{M}$ . In that case, the above formula simplify as follows:

$$I(p_\theta) = \exp(F(2\theta) - 2F(\theta)), \tag{30}$$

$$\rho(p_{\theta_1}, p_{\theta_2}) = \exp(-J_F(2\theta_1 : 2\theta_2)), \tag{31}$$

$$D_{CS}(p_{\theta_1}, p_{\theta_2}) = J_F(2\theta_1 : 2\theta_2). \tag{32}$$

The Cauchy-Schwarz divergence between mixtures of Gaussians has been reported in [35], and extended to mixtures of exponential families with conic natural parameter spaces in [33].

Furthermore, since the Jensen divergence  $J_F$  is defined for a strictly convex generator  $F$  modulo an affine term, we may choose the representative  $F(\theta) = -\log p_\theta(\omega) =: -l_\theta(\omega)$  for the equivalence class  $[F]$  of strictly convex functions, where  $\omega$  is any point belonging to the support  $\mathcal{X}$  of the exponential family  $\mathcal{X}$  and  $l_\theta(\cdot)$  the (concave) log-likelihood function, see [36] for details.

It follows that, we can rewrite the Onicescu’s informational energy, correlation coefficient and the Cauchy-Schwarz divergence when  $k(x) = 0$  as follows:

$$I(p_\theta) = \frac{p_\theta^2(\omega)}{p_{2\theta}(\omega)}, \quad \forall \omega \in \mathcal{X} \tag{33}$$

$$\rho(p_{\theta_1}, p_{\theta_2}) = \frac{\sqrt{p_{2\theta_1}(\omega)}\sqrt{p_{2\theta_2}(\omega)}}{p_{\theta_1+\theta_2}(\omega)}, \quad \forall \omega \in \mathcal{X}, \tag{34}$$

$$D_{CS}(p_{\theta_1}, p_{\theta_2}) = -\log\left(\frac{\sqrt{p_{2\theta_1}(\omega)}\sqrt{p_{2\theta_2}(\omega)}}{p_{\theta_1+\theta_2}(\omega)}\right), \tag{35}$$

$$= I_{\theta_1+\theta_2}(\omega) - \frac{I_{2\theta_1}(\omega) + I_{2\theta_2}(\omega)}{2}, \quad \forall \omega \in \mathcal{X}. \tag{36}$$

Moreover, the Cauchy-Schwarz divergence can be generalized to the broader class of Hölder divergences [28] for conjugate exponents  $\frac{1}{\alpha} + \frac{1}{\beta} = 1$  with  $\alpha > 1, \beta > 1$ , and  $\gamma > 0$  as follows:

$$D_{\text{Hölder}}^{\alpha,\gamma}(p_{\theta_1}, p_{\theta_2}) := -\log\left(\frac{\int_{\mathcal{X}} p(x)^{\gamma/\alpha} q(x)^{\gamma/\beta} d\mu(x)}{(\int_{\mathcal{X}} p(x)^\gamma d\mu(x))^{1/\alpha} (\int_{\mathcal{X}} q(x)^\gamma d\mu(x))^{1/\beta}}\right), \tag{37}$$

$$= \log\left(\frac{p_{\frac{\gamma}{\alpha}\theta_1 + \frac{\gamma}{\beta}\theta_2}(\omega)}{p_{\frac{\gamma}{\alpha}\theta_1}(\omega)p_{\frac{\gamma}{\beta}\theta_2}(\omega)}\right), \quad \forall \omega \in \mathcal{X}, \tag{38}$$

$$D_{\text{Hölder}}^{\alpha,\gamma}(p_{\lambda_1}, p_{\lambda_2}) = \log\left(\frac{p_{\frac{\gamma}{\alpha}\theta(\lambda_1) + \frac{\gamma}{\beta}\theta(\lambda_2)}(\omega)}{p_{\frac{\gamma}{\alpha}\theta(\lambda_1)}(\omega)p_{\frac{\gamma}{\beta}\theta(\lambda_2)}(\omega)}\right), \quad \forall \omega \in \mathcal{X}. \tag{39}$$

The latter two formulae hold when  $k(x) = 0$  (no auxiliary carrier term like the Gaussian family). When  $\alpha = \beta = \gamma = 2$ , we recover the Cauchy-Schwarz divergence:  $D_{\text{Hölder}}^{2,2}(p_{\theta_1}, p_{\theta_2}) = D_{CS}(p_{\theta_1}, p_{\theta_2})$ .

### 3. Some Illustrating Examples

Let us illustrate how to instantiate the generic formula with some examples of exponential families.

#### 3.1. Exponential Family of Exponential Distributions

Consider the family of exponential distributions with rate parameter  $\lambda > 0$ . The densities of this exponential family writes as  $p_\lambda(x) = \lambda \exp(-\lambda x)$  with support  $\mathcal{X} = [0, \infty)$ . We use the canonical decomposition of the exponential family to get  $t(x) = -x, \theta = \lambda, F(\theta) = -\log \theta$  and  $k(x) = 0$ . It follows that

$$I(p_\theta) = \exp(F(2\theta) - 2F(\theta)), \tag{40}$$

$$= \exp(-\log 2\theta + 2 \log \theta), \tag{41}$$

$$= \exp(-\log 2 + \log \theta), \tag{42}$$

$$= \frac{\theta}{2}. \tag{43}$$

Thus  $I(p_\lambda) = \frac{\lambda}{2}$ . Similarly, we find that

$$I(p_{\lambda_1}, p_{\lambda_2}) = \exp(F(\theta_1 + \theta_2) - F(\theta_1) - F(\theta_2)) E_{p_{\theta_1+\theta_2}}[\exp(k(x))], \tag{44}$$

$$= \frac{\lambda_1 \lambda_2}{\lambda_1 + \lambda_2}. \tag{45}$$

Thus  $\rho(p_{\lambda_1}, p_{\lambda_2}) = \frac{2\sqrt{\lambda_1\lambda_2}}{\lambda_1+\lambda_2}$ , and  $D_{CS}(p_{\lambda_1}, p_{\lambda_2}) = \log\left(\frac{\lambda_1+\lambda_2}{2}\right) - \frac{1}{2}\log(\lambda_1\lambda_2)$ . We check that  $D_{CS}(p_{\lambda_1}, p_{\lambda_2}) \geq 0$  since the arithmetic mean  $A(\lambda_1, \lambda_2) = \frac{\lambda_1+\lambda_2}{2}$  is greater or equal than the geometric mean  $G(\lambda_1, \lambda_2) = \sqrt{\lambda_1\lambda_2}$ , and  $D_{CS}(p_{\lambda_1}, p_{\lambda_2}) = \log\frac{A(\lambda_1, \lambda_2)}{G(\lambda_1, \lambda_2)}$ .

Choose  $\omega = 0$  so that  $p_\lambda(\omega) = \lambda$  and  $l_\lambda(\omega) = \log \lambda$ .

$$\rho(p_{\theta_1}, p_{\theta_2}) = \frac{\sqrt{p_{2\theta_1}(\omega)}\sqrt{p_{2\theta_2}(\omega)}}{p_{\theta_1+\theta_2}(\omega)}, \tag{46}$$

$$= \frac{2\sqrt{\lambda_1\lambda_2}}{\lambda_1 + \lambda_2}, \tag{47}$$

$$D_{CS}(p_{\theta_1}, p_{\theta_2}) = l_{\theta_1+\theta_2}(\omega) - \frac{l_{2\theta_1}(\omega) + l_{2\theta_2}(\omega)}{2}, \tag{48}$$

$$= \log(\lambda_1 + \lambda_2) - \frac{1}{2}(\log(2\lambda_1) + \log(2\lambda_2)) = \log\left(\frac{\lambda_1 + \lambda_2}{2\sqrt{\lambda_1\lambda_2}}\right). \tag{49}$$

To illustrate the fact that the formula Equation (34) is independent of the choice of  $\omega$ , let us consider  $\omega = 1$  so that  $p_\lambda(\omega) = \lambda \exp(-\lambda)$  with log-likelihood  $l_\lambda(\omega) = \log(\lambda) - \lambda$ . The correlation coefficient is then calculated as

$$\rho(p_{\theta_1}, p_{\theta_2}) = \frac{\sqrt{p_{2\theta_1}(\omega)}\sqrt{p_{2\theta_2}(\omega)}}{p_{\theta_1+\theta_2}(\omega)}, \tag{50}$$

$$= \frac{\sqrt{2\lambda_1 \exp(-2\lambda_1)}\sqrt{2\lambda_2 \exp(-2\lambda_2)}}{(\lambda_1 + \lambda_2) \exp(-(\lambda_1 + \lambda_2))}, \tag{51}$$

$$= \frac{2\sqrt{\lambda_1\lambda_2} \exp(-(\lambda_1 + \lambda_2))}{\lambda_1 + \lambda_2 \exp(-(\lambda_1 + \lambda_2))}, \tag{52}$$

$$= \frac{2\sqrt{\lambda_1\lambda_2}}{\lambda_1 + \lambda_2}. \tag{53}$$

### 3.2. Exponential Family of Poisson Distributions

The Poisson family of probability mass functions (PMFs)  $p_\lambda(x) = \frac{\lambda^x \exp(-\lambda)}{x!}$  where  $\lambda > 0$  denotes the intensity parameter and  $x \in \mathcal{X} = \{0, 1, \dots\}$  is a discrete exponential family with sufficient statistic  $t(x) = x$ , natural parameter  $\theta(\lambda) = \log \lambda$  (affine natural parameter space), cumulant function  $F(\theta) = \exp(\theta)$ , and auxiliary carrier term  $k(x) = -\log x!$ . The informational energy is

$$I(p_\lambda) = \sum_{x=0}^{\infty} p_\lambda^2(x), \tag{54}$$

$$= I(p_\theta) = \exp(F(2\theta) - 2F(\theta))E_{p_{2\theta}}[\exp(k(x))], \tag{55}$$

$$= \exp(e^{2\theta} - 2e^\theta)E_{p_{2\theta}}\left[\frac{1}{x!}\right], \tag{56}$$

$$= e^{\lambda^2 - 2\lambda}E_{p_{\lambda^2}}\left[\frac{1}{x!}\right]. \tag{57}$$

### 3.3. Exponential Family of Univariate Normal Distributions

Consider the set of univariate normal probability density function:

$$\mathcal{N} := \left\{ p_\lambda(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right), \lambda = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_{++} \right\}, \tag{58}$$

where  $\mathbb{R}_{++} = \{x \in \mathbb{R} : x > 0\}$  denote the set of positive reals. Family  $\mathcal{N}$  is interpreted as an exponential family indexed by the source parameter  $\lambda = (\mu, \sigma^2) \in \Lambda$  with  $\Lambda = \mathbb{R} \times \mathbb{R}_{++}$ .



The corresponding natural parameter is  $\theta(\lambda) = \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right)$  with the sufficient statistic  $t(x) = (x, x^2)$  on the support  $\mathcal{X} = (-\infty, \infty)$  (and no additional carrier term, i.e.,  $k(x) = 0$ ). The cumulant function for the normal family is  $F(\theta) = -\frac{\theta_1^2}{4\theta_2} + \frac{1}{2} \log\left(-\frac{\pi}{\theta_2}\right)$ .

We have

$$I(p_\theta) = \exp(F(2\theta) - 2F(\theta)), \tag{59}$$

$$= \exp\left(-\frac{(2\theta_1)^2}{4(2\theta_2)} + \frac{1}{2} \log \frac{\pi}{(-2\theta_2)} + \frac{1}{2} \frac{\theta_1^2}{\theta_2} - \log \frac{\pi}{(-\theta_2)}\right), \tag{60}$$

$$= \exp\left(\frac{1}{2} \log \frac{\pi}{(-2\theta_2)} - \log \frac{\pi}{(-\theta_2)}\right), \tag{61}$$

$$= \exp\left(-\frac{1}{2} \log(2\pi) - \frac{1}{2} \log(2\sigma^2)\right), \tag{62}$$

$$= \frac{1}{2\sigma\sqrt{\pi}}. \tag{63}$$

Similar calculations for  $\theta_1 = (\mu_1, \sigma_1)$  and  $\theta_2 = (\mu_2, \sigma_2)$  yield

$$I(p_{\mu_1, \sigma_1}, p_{\mu_2, \sigma_2}) = \frac{1}{\sqrt{2\pi}} \frac{\exp\left(-\frac{(\mu_1 - \mu_2)^2}{2\sigma_1^2 + 2\sigma_2^2}\right)}{\sqrt{\sigma_1^2 + \sigma_2^2}}. \tag{64}$$

We check that  $I(p_\theta) = I(p_\theta, p_\theta)$ .

It follows that Onicescu’s correlation coefficient between two normal densities is:

$$\rho(p_{\mu_1, \sigma_1}, p_{\mu_2, \sigma_2}) = \sqrt{\frac{2\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2}} \exp\left(-\frac{(\mu_1 - \mu_2)^2}{2\sigma_1^2 + 2\sigma_2^2}\right), \tag{65}$$

and the Cauchy-Schwarz divergence between two univariate Gaussians is:

$$D_{CS}(p_{\mu_1, \sigma_1}, p_{\mu_2, \sigma_2}) = -\log \rho(p_{\mu_1, \sigma_1}, p_{\mu_2, \sigma_2}) = \frac{(\mu_1 - \mu_2)^2}{2\sigma_1^2 + 2\sigma_2^2} + \frac{1}{2} \log\left(\frac{1}{2} \left(\frac{\sigma_1}{\sigma_2} + \frac{\sigma_2}{\sigma_1}\right)\right). \tag{66}$$

### 3.4. Exponential Family of Multivariate Normal Distributions

Consider the example of the multivariate normal (MVN) family: The parameter  $\lambda = (\lambda_v, \lambda_M)$  of a MVN consists of a vector part  $\lambda_v = \mu$  and a  $d \times d$  positive-definite matrix part  $\lambda_M = \Sigma \succ 0$ . The density is given by

$$p_\lambda(x; \lambda) = \frac{1}{(2\pi)^{\frac{d}{2}} \sqrt{|\lambda_M|}} \exp\left(-\frac{1}{2} (x - \lambda_v)^\top \lambda_M^{-1} (x - \lambda_v)\right), \tag{67}$$

where  $|\cdot|$  denotes the matrix determinant. Choose the sufficient statistic  $t(x) = (x, -\frac{1}{2}xx^\top)$  so that  $\theta = (\theta_v = \Sigma^{-1}\mu, \theta_M = \Sigma^{-1})$ . Since  $k(x) = 0$ , let  $\omega = 0$ ,

$$p_\lambda(0) = \frac{1}{(2\pi)^{\frac{d}{2}} \sqrt{|\Sigma|}} \exp\left(-\frac{1}{2} \mu^\top \Sigma^{-1} \mu\right),$$

and apply the formula of Equation (33) with  $2\theta_M = 2\Sigma^{-1} = \left(\frac{1}{2}\Sigma\right)^{-1}$ :

$$I(p_\theta) = \frac{p_\theta^2(0)}{p_{2\theta}(0)}, \tag{68}$$

$$= \frac{(2\pi)^{\frac{d}{2}} \left|\frac{1}{2}\Sigma\right|^{\frac{1}{2}}}{(2\pi)^d |\Sigma|} = \frac{1}{2^d \pi^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}}. \tag{69}$$

Let us calculate the formula for the Cauchy-Schwarz divergence between two multivariate Gaussian distributions. We have  $\theta(\lambda) = (\theta_b, \theta_M) = (\Sigma^{-1}\mu, \Sigma^{-1})$  for  $\lambda = (\mu, \Sigma)$ . Conversely, we have  $\lambda(\theta) = (\theta_M^{-1}\theta_b, \theta_M^{-1})$ . It follows that

$$\lambda(\theta_1 + \theta_2) = \left( (\Sigma_1^{-1} + \Sigma_2^{-1})^{-1} (\Sigma_1^{-1}\mu_1 + \Sigma_2^{-1}\mu_2), (\Sigma_1^{-1} + \Sigma_2^{-1})^{-1} \right). \tag{70}$$

In particular, we have  $\lambda(2\theta) = (\mu, \frac{1}{2}\Sigma)$ . Let  $\omega = 0$  so that

$$p_\lambda(0) = \frac{1}{(2\pi)^{\frac{d}{2}} \sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}\mu^\top \Sigma^{-1}\mu\right), \tag{71}$$

$$l_\lambda(0) = -\frac{d}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma| - \frac{1}{2}\mu^\top \Sigma^{-1}\mu. \tag{72}$$

Thus we get

$$D_{CS}(p_{\lambda_1}, p_{\lambda_2}) = l_{\lambda(\theta_1+\theta_2)}(\omega) - \frac{l_{\lambda(2\theta_1)}(\omega) + l_{\lambda(2\theta_2)}(\omega)}{2}, \tag{73}$$

$$\begin{aligned} D_{CS}(p_{\mu_1, \Sigma_1}, p_{\mu_2, \Sigma_2}) &= \frac{1}{2} \log \left( \frac{1}{2^d} \frac{\sqrt{|\Sigma_1| |\Sigma_2|}}{|\Sigma_1^{-1} + \Sigma_2^{-1}|} \right) \\ &\quad + \frac{1}{2} \mu_1^\top \Sigma_1^{-1} \mu_1 + \frac{1}{2} \mu_2^\top \Sigma_2^{-1} \mu_2 \\ &\quad - \frac{1}{2} (\Sigma_1^{-1} \mu_1 + \Sigma_2^{-1} \mu_2)^\top (\Sigma_1^{-1} + \Sigma_2^{-1})^{-1} (\Sigma_1^{-1} \mu_1 + \Sigma_2^{-1} \mu_2). \end{aligned} \tag{74}$$

The formula coincides with the formula of Equation (66) when  $\Sigma_1 = \sigma_1^2$  and  $\Sigma_2 = \sigma_2^2$ .

### 3.5. Exponential Family of Pareto Distributions

Consider the family of Pareto densities defined by a shape parameter  $a > 0$  and a prescribed scale parameter  $k > 0$  as follows:

$$\left\{ p_a(x) = \frac{ak^a}{x^{a+1}}, \quad x \in [k, \infty) \right\}. \tag{75}$$

Writing the density as  $p_a(x) = \exp(a \log k + \log a - (a + 1) \log x) = p_{\theta(a)}(x)$ , we deduce that the Pareto densities form an exponential family with natural parameter  $\theta = a + 1$ , sufficient statistic  $t(x) = \log x$  and  $k(x) = 0$ . Let us choose  $\omega = k$ , and apply the generic formula for the informational energy with  $\theta(a) = a + 1$  and  $\theta^{-1}(b) = b - 1$  (and  $2\theta(a) = \theta^{-1}(2a + 2) = 2a + 2 - 1 = 2a + 1$ ):

$$I(p_\theta) = \frac{p_\theta^2(\omega)}{p_{2\theta}(\omega)}, \tag{76}$$

$$I(p_a) = \frac{p_{\theta(a)}^2(k)}{p_{2\theta(a)}(k)} = \frac{p_a^2(k)}{p_{2a+1}(k)} \tag{77}$$

$$= \left(\frac{ak^a}{x^{a+1}}\right)^2 \frac{k^{2a+2}}{(2a+1)k^{2a+1}}, \tag{78}$$

$$= \frac{a^2}{k(2a+1)}. \tag{79}$$

### 3.6. Instantiating Formula with a Computer Algebra System

In general, we can automate the calculations of closed-form formula for conic exponential families using a computer algebra system (CAS) by defining the source-to-natural parameter conversion function  $\theta(\lambda)$ , and then apply the formula

$$D_{CS}(p_{\lambda_1}, p_{\lambda_2}) = \log \left( \frac{p_{\theta(\lambda_1)+\theta(\lambda_2)}(\omega)}{\sqrt{p_{2\theta(\lambda_1)}(\omega)}\sqrt{p_{2\theta(\lambda_2)}(\omega)}} \right), \quad \forall \omega \in \mathcal{X}. \tag{80}$$

For example, using the CAS Maxima (<http://maxima.sourceforge.net/> accessed on 14 April 2022), we can calculate the formula of the information energy of Pareto densities as follows:

```
/* Pareto densities form an exponential family */
assume(k>0);
assume(a>0);

Pareto(x, a) := a*(k**a)/(x**(a+1));

/* check that it is a density (=1) */
integrate(Pareto(x, a), x, k, inf);

/* calculate Onicescu's informational energy */
integrate(Pareto(x, a)**2, x, k, inf);

/* method bypassing the integral calculation */
omega:k;
(Pareto(omega, a)**2)/Pareto(omega, 2*a+1);
```

### 4. Informational Energy and the Laws of Thermodynamics

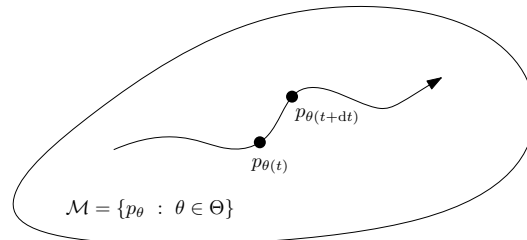
The informational energy was originally motivated by an analogy to kinetic energy in physics, and proves useful when investigating the thermodynamics laws on a statistical manifold [20] where thermodynamic processes [37] can be viewed as oriented trajectories on the manifold as depicted in Figure 1: Indeed, the third law of thermodynamics states that the entropy and the kinetic energy (i.e., informational energy) of an isolated thermodynamical system should vary with opposite directions. Thus when viewing a parametric family  $\mathcal{M} = \{p_\theta : \theta \in \Theta\}$  of distributions as a statistical manifold [20,38], a valid oriented thermodynamic path  $\{p_{\theta(t)}, t \in I \subset \mathbb{R}\}$  (with time  $t$  increasing) should satisfy the following condition:

$$\left( H(p_{\theta(t+dt)}) - H(p_{\theta(t)}) \right) \times \left( I(p_{\theta(t+dt)}) - I(p_{\theta(t)}) \right) < 0.$$

This condition can be written equivalently as the following variational inequality:

$$\frac{d}{dt}H(p_{\theta(t)}) \times \frac{d}{dt}I(p_{\theta(t)}) < 0. \tag{81}$$

We consider the thermodynamic process paths on exponential family manifolds in Section 4.1 and on location-scale manifolds in Section 4.2.



**Figure 1.** Visualizing a thermodynamic process as an oriented trajectory on a statistical manifold.

#### 4.1. Exponential Family Manifolds

Consider statistical manifolds induced by exponential families [20], and let us report the variations of entropy and informational energy on those manifolds.

**Lemma 1.** *The variation of entropy of a density  $p_{\theta(t)}$  of a  $D$ -dimensional exponential family with log-normalized  $F(\theta)$  for  $\theta = (\theta_1, \dots, \theta_D)$  is*

$$\frac{d}{dt}H(p_{\theta(t)}) = -\dot{\theta}(t)^\top \nabla^2 F(\theta(t)) \theta(t),$$

where  $\dot{\theta}(t) = \left(\frac{d}{dt}\theta_1(t), \dots, \frac{d}{dt}\theta_D(t)\right)$ .

**Proof.** The entropy  $H(p_{\theta})$  of a density  $p_{\theta}$  of a (potentially reparametrized) natural exponential family with log-normalizer  $F(\theta)$  is given by [34]:

$$H(p_{\theta}) = -F^*(\eta) = F(\theta) - \langle \theta, \nabla F(\theta) \rangle,$$

where  $\langle \cdot, \cdot \rangle$  denotes the scalar product (Euclidean inner product). It follows that we have

$$\begin{aligned} \frac{d}{dt}H(p_{\theta(t)}) &= \frac{d}{dt} (F(\theta(t)) - \langle \theta(t), \nabla F(\theta(t)) \rangle), \\ &= \langle \dot{\theta}(t), \nabla F(\theta(t)) \rangle - \langle \dot{\theta}(t), \nabla F(\theta(t)) \rangle - \langle \theta(t), \langle \nabla^2 F(\theta(t)), \dot{\theta}(t) \rangle \rangle, \\ &= -\dot{\theta}(t)^\top \nabla^2 F(\theta(t)) \theta(t). \end{aligned}$$

Notice that the Hessian matrix  $\nabla^2 F(\theta)$  is positive-definite since the log-normalizer  $F(\theta)$  is strictly convex for a minimal regular exponential family [20].  $\square$

Next, we report the variation of the informational energy on any thermodynamic process trajectory:

**Lemma 2.** *The variation of informational energy of a density  $p_{\theta(t)}$  of an exponential family of order  $D$  with log-normalizer  $F(\theta)$  and zero auxiliary carrier term  $k(x)$  is*

$$\frac{d}{dt}I(p_{\theta(t)}) = 2 \langle \dot{\theta}(t), \nabla F(2\theta(t)) - \nabla F(\theta(t)) \rangle I(p_{\theta(t)}),$$

provided that  $2\theta \in \Theta$ .

**Proof.** Assume no extra auxiliary carrier term (i.e.,  $k(x) = 0$ ) so that by Theorem 1 we have  $I(p_\theta) = \exp(F(2\theta) - 2F(\theta))$ . Therefore, we get

$$\frac{d}{dt}I(p_\theta(t)) = 2\langle \dot{\theta}(t), \nabla F(2\theta(t)) - \nabla F(\theta(t)) \rangle I(p_\theta(t)).$$

□

Thus we can check that whether the statistical normal manifold [20] satisfies the third thermodynamic law everywhere or not as follows: We have  $H(p_{\mu(t),\sigma(t)}) = \log(\sqrt{2\pi e}\sigma(t))$  and  $I(p_{\mu(t),\sigma(t)}) = \frac{1}{2\sqrt{\pi}\sigma(t)}$ . Therefore we get:

$$\begin{aligned} \frac{d}{dt}H(p_{\theta(t)}) \times \frac{d}{dt}I(p_{\theta(t)}) &= \frac{\dot{\sigma}(t)}{\sigma(t)} \times \left( -\dot{\sigma}(t) \frac{1}{2\sqrt{\pi}} \sigma^2(t) \right), \\ &= -\frac{1}{2\sqrt{\pi}} \frac{\dot{\sigma}(t)^2}{\sigma^3(t)} < 0, \end{aligned}$$

since  $\sigma(t) > 0$  for all  $t$ . Any smooth curve on a statistical normal manifold can be a thermodynamic process path satisfying the third law of thermodynamics.

#### 4.2. Location-Scale Manifolds

Consider the setting where  $\mathcal{M}$  is a statistical manifold modeling a location-scale family:

$$\mathcal{M} = \left\{ p_{l,s}(x) = \frac{1}{s} p_{\text{std}}\left(\frac{x-l}{s}\right), \quad (l, s) \in (\mathbb{R}, \mathbb{R}_{++}) \right\},$$

where  $p_{\text{std}}(x)$  denotes the standard probability density function of the family (i.e.,  $p_{\text{std}}(x) = p_{0,1}(x)$ ). The family of normal distributions and the family of Cauchy distributions are two examples of location-scale families with standard density  $p_{\text{std}}(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right)$  and  $p_{\text{std}}(x) = \frac{1}{\pi(1+x^2)}$ , respectively.

By a change of variable  $y = \frac{x-l}{s}$  in the integral definitions of the entropy (Equation (4)) and the informational energy (Equation (1)) of a location-scale probability density function  $p_{l,s}(x)$ , we find that

$$H(p_{l,s}) = H(p_{\text{std}}) + \log s, \tag{82}$$

$$I(p_{l,s}) = \frac{1}{s} I(p_{\text{std}}). \tag{83}$$

The informational energy of the standard normal and Cauchy distributions are  $\frac{1}{2\sqrt{\pi}}$  (since  $\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}$ ) and  $\frac{1}{2}$  (since  $\int_{-\infty}^{\infty} \frac{1}{(1+x^2)^2} dx = \frac{\pi}{2}$ ), respectively. Thus the informational energy of the normal distributions  $N(\mu, \sigma)$  and the Cauchy distributions  $C(l, s)$  are  $\frac{1}{2\sigma\sqrt{\pi}}$  (see also Table 2) and  $\frac{1}{2s}$ , respectively.

Using a first-order Taylor expansion on the scale parameter

$$s(t + dt) \approx s(t) + \dot{s}(t)dt,$$

we get

$$\frac{d}{dt}H(p_{l(t),s(t)}) = \dot{s}(t),$$

and

$$\frac{d}{dt}I(p_{l(t),s(t)}) = -\frac{\dot{s}(t)}{s^2(t)} I(p_{\text{std}}).$$

Therefore, we have

$$\frac{d}{dt}H(p_{\theta(t)}) \times \frac{d}{dt}I(p_{\theta(t)}) = -\frac{\dot{s}(t)^2}{s^3(t)}I(p_{\text{std}}) < 0,$$

since  $s(t) > 0$  and  $I(p_{\text{std}}) > 0$  for any (standard) probability density function. Thus all smooth paths on a location-scale manifold are compatible with the third law of thermodynamics.

Note that the second law of thermodynamics states that the entropy of an isolated system increases over time. That is, for a thermodynamic process  $\{p_{\theta(t)}, t \in I \subset \mathbb{R}\}$ , we shall have  $H(p_{\theta(t)}) < H(p_{\theta(t+dt)})$ . On a statistical location-scale manifold with differential entropy  $H(p_{l(t),s(t)}) = H(p_{\text{std}}) + \log s(t)$ , we have

$$\frac{d}{dt}H(p_{l(t),s(t)}) = \frac{\dot{s}(t)}{s(t)}.$$

Thus the entropy of a thermodynamic process increases whenever the scale  $s(t)$  increases with time (i.e.,  $\dot{s}(t) > 0$ ) since  $s(t) > 0$ . Therefore not all oriented thermodynamic process paths on a location-scale manifold satisfies the second law of thermodynamics but only paths with increasing scales.

## 5. Summary and Discussion

Shannon's entropy [11] and Onicescu's informational energy [5] are two complementary measures of uncertainty of a random variable: When randomness increases, Shannon's entropy increases but Onicescu's informational energy decreases. Table 1 compares the properties of these randomness measures: Interestingly, the Onicescu's informational energy is strictly convex and is always positive while Shannon's entropy is strictly concave and can be negative for continuous random variables (e.g., differential entropy of a normal distribution with small variance). The fundamental quantity of the informational energy of Equation (1) has been studied in various other fields under different names: Simpson's diversity index [16] or Gini-Simpson index [18] in ecology, Herfindahl–Hirschman index [13] in economics, or the index of coincidence [14] in cryptanalysis. In this work, we show how to compute the informational energy for a distribution belonging to an exponential family in Theorem 1, and reported closed-form formula for its associated correlation coefficient and related statistical divergences. We then show how to characterize thermodynamic processes satisfying the third law of thermodynamics by interpreting them as smooth oriented trajectory paths on a statistical manifold and asserting that the entropy and the informational energy (interpreted as the kinetic energy) shall vary in opposite directions.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

**Acknowledgments:** The author would like to thank the reviewers for their comments and suggestions which lead to this revised manuscript.

**Conflicts of Interest:** The author declares no conflict of interest.

## References

1. Billingsley, P. *Probability and Measure*, 3rd ed.; Wiley: Hoboken, NJ, USA, 1995.
2. Iosifescu, M. Obituary notice: Octav Onicescu, 1892–1983. *Int. Stat. Rev.* **1986**, *54*, 97–108.
3. Crepel, P.; Fienberg, S.; Gani, J. *Statisticians of the Centuries*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2013.
4. Onicescu, O.; Stefanescu, V. *Elements of Informational Statistics with Applications/Elemente de Statistica Informationala cu Aplicatii (in Romanian)*; Editura Tehnica: Bucharest, Romania, 1979.
5. Onicescu, O. Théorie de l'information énergie informationelle. *Comptes Rendus De L'Academie Des Sci. Ser. AB* **1966**, *263*, 841–842.

6. Pardo, L.; Taneja, I. Information Energy and Its Applications. In *Advances in Electronics and Electron Physics*; Elsevier: Amsterdam, The Netherlands, 1991; Volume 80, pp. 165–241.
7. Pardo, J.A.; Vicente, M. Asymptotic distribution of the useful informational energy. *Kybernetika* **1994**, *30*, 87–99.
8. Nielsen, F.; Nock, R. On Rényi and Tsallis entropies and divergences for exponential families. *arXiv* **2011**, arXiv:1105.3259.
9. Vajda, I.; Zvárová, J. On generalized entropies, Bayesian decisions and statistical diversity. *Kybernetika* **2007**, *43*, 675–696.
10. Ho, S.W.; Yeung, R.W. On the discontinuity of the Shannon information measures. *IEEE Trans. Inf. Theory* **2009**, *55*, 5362–5374.
11. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [[CrossRef](#)]
12. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; John Wiley & Sons: Hoboken, NJ, USA, 2012.
13. Matsumoto, A.; Merlone, U.; Szidarovszky, F. Some notes on applying the Herfindahl–Hirschman Index. *Appl. Econ. Lett.* **2012**, *19*, 181–184. [[CrossRef](#)]
14. Harremoës, P.; Topsøe, F. Inequalities between entropy and index of coincidence derived from information diagrams. *IEEE Trans. Inf. Theory* **2001**, *47*, 2944–2960. [[CrossRef](#)]
15. Friedman, W.F. *The Index of Coincidence and Its Applications in Cryptography*; Riverbank Laboratories: Geneva, IL, USA, 1922; Department of Ciphers Publ 22.
16. Simpson, E.H. Measurement of diversity. *Nature* **1949**, *163*, 688–688. [[CrossRef](#)]
17. Nunes, A.P.; Silva, A.C.; Paiva, A.C.D. Detection of masses in mammographic images using geometry, Simpson’s Diversity Index and SVM. *Int. J. Signal Imaging Syst. Eng.* **2010**, *3*, 40–51. [[CrossRef](#)]
18. Rao, C.R. Gini-Simpson index of diversity: A characterization, generalization and applications. *Util. Math.* **1982**, *21*, 273–282.
19. Nielsen, F.; Boltz, S. The Burbea-Rao and Bhattacharyya centroids. *IEEE Trans. Inf. Theory* **2011**, *57*, 5455–5466. [[CrossRef](#)]
20. Calin, O.; Udriște, C. *Geometric Modeling in Probability and Statistics*; Springer: Berlin/Heidelberg, Germany, 2014.
21. Agop, M.; Gavriluț, A.; Rezuș, E. Implications of Onicescu’s informational energy in some fundamental physical models. *Int. J. Mod. Phys. B* **2015**, *29*, 1550045. [[CrossRef](#)]
22. Chatzisavvas, K.C.; Moustakidis, C.C.; Panos, C. Information entropy, information distances, and complexity in atoms. *J. Chem. Phys.* **2005**, *123*, 174111. [[CrossRef](#)] [[PubMed](#)]
23. Alipour, M.; Mohajeri, A. Onicescu information energy in terms of Shannon entropy and Fisher information densities. *Mol. Phys.* **2012**, *110*, 403–405. [[CrossRef](#)]
24. Ou, J.H.; Ho, Y.K. Shannon, Rényi, Tsallis Entropies and Onicescu Information Energy for Low-Lying Singly Excited States of Helium. *Atoms* **2019**, *7*, 70. [[CrossRef](#)]
25. Andonie, R.; Cataron, A. An information energy LVQ approach for feature ranking. In Proceedings of the European Symposium on Artificial Neural Networks, Bruges, Belgium, 28–30 April 2004.
26. Rizescu, D.; Avram, V. Using Onicescu’s informational energy to approximate social entropy. *Procedia Soc. Behav. Sci.* **2014**, *114*, 377–381. [[CrossRef](#)]
27. Jenssen, R.; Principe, J.C.; Erdogmus, D.; Eltoft, T. The Cauchy–Schwarz divergence and Parzen windowing: Connections to graph theory and Mercer kernels. *J. Frankl. Inst.* **2006**, *343*, 614–629. [[CrossRef](#)]
28. Nielsen, F.; Sun, K.; Marchand-Maillet, S. On Hölder projective divergences. *Entropy* **2017**, *19*, 122. [[CrossRef](#)]
29. Nielsen, F.; Garcia, V. Statistical exponential families: A digest with flash cards. *arXiv* **2009**, arXiv:0911.4863.
30. Barndorff-Nielsen, O. *Information and Exponential Families*; John Wiley & Sons: Hoboken, NJ, USA, 2014.
31. Nielsen, F.; Nock, R. A closed-form expression for the Sharma–Mittal entropy of exponential families. *J. Phys. A: Math. Theor.* **2011**, *45*, 032003. [[CrossRef](#)]
32. Nielsen, F.; Nock, R. On the chi square and higher-order chi distances for approximating f-divergences. *IEEE Signal Process. Lett.* **2013**, *21*, 10–13. [[CrossRef](#)]
33. Nielsen, F. Closed-form information-theoretic divergences for statistical mixtures. In Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), Tsukuba, Japan, 11–15 November 2012 ; pp. 1723–1726.
34. Nielsen, F.; Nock, R. Entropies and cross-entropies of exponential families. In Proceedings of the 2010 IEEE International Conference on Image Processing, Hong Kong, China, 26–29 September 2010; pp. 3621–3624.
35. Kampa, K.; Hasanbelliu, E.; Principe, J.C. Closed-form Cauchy-Schwarz PDF divergence for mixture of Gaussians. In Proceedings of the International Joint Conference on Neural Networks, San Jose, CA, USA, 31 July–5 August 2011; pp. 2578–2585.
36. Nielsen, F.; Nock, R. Cumulant-free closed-form formulas for some common (dis)similarities between densities of an exponential family. *arXiv* **2020**, arXiv:2003.02469.
37. Ito, S.; Dechant, A. Stochastic time evolution, information geometry, and the Cramér-Rao bound. *Phys. Rev. X* **2020**, *10*, 021056. [[CrossRef](#)]
38. Nielsen, F. The Many Faces of Information Geometry. *Not. Am. Math. Soc.* **2022**, *69*, 36–45. [[CrossRef](#)]