

## Article

# Skewed Jensen—Fisher Divergence and Its Bounds

Takuya Yamano 

Faculty of Science, Kanagawa University, 2946, 6-233 Tsuchiya, Kanagawa, Hiratsuka 259-1293, Japan; yamano@amy.hi-ho.ne.jp

**Abstract:** A non-uniform (skewed) mixture of probability density functions occurs in various disciplines. One needs a measure of similarity to the respective constituents and its bounds. We introduce a skewed Jensen–Fisher divergence based on relative Fisher information, and provide some bounds in terms of the skewed Jensen–Shannon divergence and of the variational distance. The defined measure coincides with the definition from the skewed Jensen–Shannon divergence via the de Bruijn identity. Our results follow from applying the logarithmic Sobolev inequality and Poincaré inequality.

**Keywords:** skewed Jensen–Fisher divergence; relative Fisher information; Fisher information; logarithmic Sobolev inequality; Poincaré inequality

## 1. Introduction

Comparison of a probability density function with a mixture, i.e., a weighted sum of several density functions, is frequently needed in various disciplines, such as statistics, information theory, signal processing, bioinformatics, machine learning, neuroscience, natural language processing, time series analysis, and many others. To quantify the closeness of the mixed density function to the respective one, the Jensen–Shannon divergence (JSD) has been used since the definition by Lin [1]. JSD is a symmetrized Kullback–Leibler (KL) divergence with an equally weighted mixture. The skewed version of JSD has also been introduced as the generalized JSD in [2,3]. A symmetrized JSD through a skewed KL divergence has been studied in [2,3]. By the skewed divergence, we mean that one chooses its reference density function as an unequally weighted density.

Recently, on the other hand, a Jensen–Fisher divergence (JFD) has been introduced [4] to properly detect the highly oscillatory behavior of density functions. The JFD uses relative Fisher information instead of a KL divergence, by which it is more sensitive to the change of densities because of the gradient feature, whereas the KL divergence encapsulates the overall feature. In this paper, we show that the skewed version of the Jensen–Fisher divergence can be defined similarly to the skewed JSD, and express it in terms of the Fisher information of respective density functions.

In the next two sections, we define it in two different ways, but we see that they are identical. Some backgrounds for relative Fisher information and the previous applications to physical sciences are also mentioned there. Then, we provide some lower bounds by applying some integral inequalities for gradient functions. In the last section, we remark on the application of the Sobolev inequality to the skewed Jensen–Fisher divergence.

## 2. Definition via Jensen–Shannon Divergence

Let  $p(x)$  and  $q(x)$  be two continuous probability density functions on  $\mathbf{R}^n$ . The skewed Jensen–Shannon divergence is defined as the weighted sum of the respective skewed KL divergence:

$$JSD_{\alpha}(p, q) := (1 - \alpha)KL(p \parallel (1 - \alpha)p + \alpha q) + \alpha KL(q \parallel (1 - \alpha)p + \alpha q), \quad (1)$$



**Citation:** Yamano, T. Skewed Jensen—Fisher Divergence and Its Bounds. *Foundations* **2021**, *1*, 256–264. <https://doi.org/10.3390/foundations1020018>

Academic Editor: Ioannis G. Tollis

Received: 21 October 2021

Accepted: 11 November 2021

Published: 16 November 2021

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



**Copyright:** © 2021 by the author. 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/>).

where  $\alpha \in [0, 1]$  controls the weights in the mixture. We will use the separation symbol  $\|$  between the two functions for the single divergence and the comma , for the combined divergences. Here, the skewed KL divergence of order  $\alpha$  from  $p(x)$  to  $q(x)$  is defined as [1,5,6]

$$KL(p\|(1-\alpha)p + \alpha q) := \int_{\mathbb{R}^n} p \log \frac{p}{(1-\alpha)p + \alpha q} dx. \quad (2)$$

We always assume absolute continuity when we encounter a division of density functions. Similarly, from  $q(x)$  to  $p(x)$

$$KL(q\|(1-\alpha)p + \alpha q) := \int_{\mathbb{R}^n} q \log \frac{q}{(1-\alpha)p + \alpha q} dx. \quad (3)$$

Thus, the weighted sum of the above two skewed KL divergences of order  $\alpha$  provides the following expression for the definition Equation (1)

$$JSD_{\alpha}(p, q) = S((1-\alpha)p + \alpha q) - (1-\alpha)S(p) - \alpha S(q), \quad (4)$$

where  $S(p) = -\int p \log p dx$  is the differential entropy of an arbitrary random variable with a density function  $p(x)$ , and others are defined similarly. For  $p \neq q$ ,  $JSD_{\alpha}(p, q)$  shows positivity, which means equivalently that  $S(p)$  is a concave function. When  $\alpha = 1/2$ , Equation (1) gives the usual Jensen–Shannon divergence (JSD) [1,7,8]:

$$\begin{aligned} JSD(p, q) &= \frac{1}{2} \int_{\mathbb{R}^n} p \log \frac{2p}{p+q} dx + \frac{1}{2} \int_{\mathbb{R}^n} q \log \frac{2q}{p+q} dx \\ &= S\left(\frac{p+q}{2}\right) - \frac{1}{2}(S(p) + S(q)), \end{aligned} \quad (5)$$

which is a symmetric measure with respect to  $p$  and  $q$  and quantifies the deviation of the mean from their respective density functions in terms of the differential entropy.

**Remark 1.** A special instance of  $\alpha = 1/2$  in Equation (2) was considered in [9]. The definition Equation (1) is also used in [10]. Lin [1] introduced the weighted combination of any finite number of density functions into the Jensen–Shannon divergence in terms of the differential entropy of the constituent functions. He termed it the generalized Jensen–Shannon divergence. However, the expression Equation (1), in terms of the skewed KL divergences, is useful as it clearly shows a distance of the mixture from the respective density functions, and this form motivates us to introduce the definition of the skewed Jensen–Fisher divergence in a similar way as we will see below. A special case of equal weight in the above definition has previously been defined in [2,3]:

$$\frac{1}{2}KL(p\|(1-\alpha)p + \alpha q) + \frac{1}{2}KL(q\|(1-\alpha)p + \alpha q).$$

To proceed, recall that the de Bruijn identity [11–13] relates the differential entropy of  $p$  to its Fisher information and it has been extended in such a way that the detailed statistics of the noise do not enter [14]:

$$\left. \frac{d}{d\delta} S(p + \sqrt{\delta}\eta) \right|_{\delta=0} = \frac{1}{2}I(p), \quad (6)$$

where  $\eta$  is an arbitrary symmetric density function with zero mean and unit variance, which is not necessarily a standard Gaussian. Fisher information of  $p(x)$  on  $\mathbb{R}^n$  is defined [15,16] as

$$I(p) := \int_{\mathbb{R}^n} \frac{\|\nabla p\|^2}{p} dx, \quad (7)$$

where, as usual,  $\|\cdot\|$  denotes  $l_2$ -norm on  $\mathbb{R}^n$ . Fisher information, in general, reflects the gradient content (i.e., sharpness) of probability density functions. This feature is informative as a measure because it is more sensitive to the degree of localization and oscillatory behavior than using differential entropy. Indeed, it has been used to detect information on the radial densities of relativistic and non-relativistic hydrogenic atoms (e.g., [17,18], and many others).

Note that, for the perturbed density functions  $p_\delta = p + \sqrt{\delta}\eta$  and  $q_\delta = q + \sqrt{\delta}\eta$ , we have

$$\begin{aligned} \frac{d}{d\delta} S((1-\alpha)p_\delta + \alpha q_\delta) \Big|_{\delta=0} &= \frac{d}{d\delta} S((1-\alpha)p + \alpha q + \sqrt{\delta}\eta) \Big|_{\delta=0} \\ &= \frac{1}{2} I((1-\alpha)p + \alpha q). \end{aligned} \quad (8)$$

Thus, in view of Equation (4), the skewed Jensen–Fisher divergence is induced by the skewed Jensen–Shannon divergence, as described in the following result:

**Proposition 1.** The derivative of the skewed Jensen–Shannon divergence between the two perturbed density functions  $p_\delta$  and  $q_\delta$  can be given as half of the skewed Jensen–Fisher divergence, with the same order between the two original density functions:

$$\frac{d}{d\delta} JSD_\alpha(p_\delta, q_\delta) \Big|_{\delta=0} = -\frac{1}{2} JFD_\alpha(p, q), \quad (9)$$

by which we can define the skewed Jensen–Fisher divergence of order  $\alpha$  as follows:

**Definition 1.** For  $\alpha \in [0, 1]$ , let  $p$  and  $q$  be two probability density functions. Then, it is defined by

$$JFD_\alpha(p, q) = (1-\alpha)I(p) + \alpha I(q) - I((1-\alpha)p + \alpha q). \quad (10)$$

This definition is consistent with another one given in the next section. The special case of  $\alpha = 1/2$  was devised and applied to see the behavior of some probability distributions in [4]:

$$\frac{d}{d\delta} JSD_{\frac{1}{2}}(p, q) \Big|_{\delta=0} = \frac{1}{2} I\left(\frac{p+q}{2}\right) - \frac{1}{4}(I(p) + I(q)) = -\frac{1}{2} JFD_{\frac{1}{2}}(p, q). \quad (11)$$

The defined form Equation (10) can be advantageous, because we can control the mixture of Fisher information by changing the skew parameter  $\alpha$ .

**Remark 2.** A connection to nonequilibrium processes through the de Bruijn-type identity can be obtained. The probability density functions  $p(x, t)$  and  $q(x, t)$  for systems described by heat equations and linear Fokker–Planck equations satisfy the de Bruijn-type identity [19]

$$\frac{d}{dt} KL(p||q) = -\mathcal{D}I(p|q),$$

where  $t$  is time and  $\mathcal{D}$  is the diffusion constant associated with the probability current of the equations. We will define  $I(p|q)$  in the right-hand side (i.e., the relative Fisher information)

soon in the next section. Thus, we have an expression of  $JFD_\alpha(p, q)$  in terms of the time derivative of the skewed KL divergence:

$$JFD_\alpha(p, q) = -\frac{1}{\mathcal{D}} \frac{d}{dt} \{KL(p \parallel (1 - \alpha)p + \alpha q) + KL(q \parallel (1 - \alpha)p + \alpha q)\}. \quad (12)$$

### 3. Definition via Relative Fisher Information

The relative Fisher information of  $p$  with respect to  $q$  is defined as

$$I(p|q) := \int_{\mathbf{R}^n} p(x) \left\| \nabla \log \frac{p(x)}{q(x)} \right\|^2 dx, \quad (13)$$

where we use the separation symbol  $|$  between  $p$  and  $q$  instead of  $\parallel$ , according to the convention. It is an asymmetric measure in  $p$  and  $q$ . It is non-negative and vanishes when  $p = q$ , which is desirable as a measure of statistical distance. Moreover, one does not suffer from a difficult computation of the normalization factor of  $q$ , which is often non-normalized in many cases: that is,  $I(p|kq) = I(p|q)$  for any constant  $k > 0$ . This merit has recently been used in the machine learning community, and in related fields such as Bayesian statistical inference [20–25]. Since this measure involves derivatives of density functions, it is more sensitive to reflect the oscillatory behavior in density functions that appear in quantum mechanical systems [4]. That is, since the probability densities of quantum systems such as a particle in a quantum potential well, isotropic harmonic oscillators, and hydrogen-like atoms oscillate spatially, and these characters increase as the corresponding quantum numbers increase, the Jensen–Shannon divergence is not necessarily an informative measure.

#### Some Background on Relative Fisher Information

First, we mention the original source of relative Fisher information, which has been defined and was named differently in different disciplines. As far as the author knows, the form was first proposed in 1978 in Equations (14) and (25) by Hammad in [26], in which the gradient of a density function  $f(x, \theta)$  is taken with respect to the parameter  $\theta$ :

$$\int_{\mathbf{R}^n} p(x, \theta) \left( \frac{\partial}{\partial \theta} \log \frac{p(x, \theta)}{q(x, \theta)} \right)^2 dx. \quad (14)$$

In 1986, Barron [27] defined the counterpart of the above by replacing the  $\theta$ -derivative with the  $x$ -derivative (i.e., the so-called shift transformation families) to show that the sequence of Fisher information converges to the Cramér–Rao bound. He termed it *standardized Fisher information*. Johnson [28,29] in 2000 has also defined the same form as Equation (13) in connection with the central limit theorem and termed it as *Fisher information distance*. Moreover, in mathematics, Otto and Villani introduced it with a measure theoretical notation as an equivalent of the logarithmic Sobolev inequality [30,31]. On the other hand, Hyvärinen [20,21], in 2005, was able to reach the proposal of the same form to estimate statistical models whose normalization constant of a density function is not known in closed form. In a recent machine learning community, some researchers call it Fisher divergence [22,23]. However, Hammad used the term Fisher divergence to denote the symmetrized form  $I(p|q) + I(q|p)$  (see Equation (41) of [26]). Relative Fisher information is potentially useful in many areas of science. In quantum mechanical applications, the excited states of density functions (the square of the wave functions) are compared with its ground state one, and the information shows significant changes according to the eigen states in respective quantum systems under potentials (e.g., [17,32–38]). We mention two more examples of the role of relative Fisher information in statistical physics. It connects the phase space gradient of the dissipated work in nonequilibrium processes: the distance between canonical equilibrium densities that correspond to forward and backward processes is proportional to the average of the squared gradient of the work dissipated

into the environment [39]. When heat exchange occurs between two bodies with different temperatures under the exchange fluctuation theorem, relative Fisher information of the heat probability with respect to its reverse is proportional to the square of the inverse temperature difference before contact [40].

We now define the skewed Jensen–Fisher divergence of order  $\alpha$  as follows:

**Definition 2.** For  $\alpha \in [0, 1]$ , let  $p$  and  $q$  be two probability density functions. Then it is defined by

$$JFD_{\alpha}(p, q) = (1 - \alpha)I(p|(1 - \alpha)p + \alpha q) + \alpha I(q|(1 - \alpha)p + \alpha q). \quad (15)$$

When  $\alpha = 0$  and  $\alpha = 1$ , it vanishes. By this definition and Equation (13), the skewed Jensen–Fisher divergence of order  $\alpha$  can be expressed by the Fisher information of respective density functions

$$\begin{aligned} JFD_{\alpha}(p, q) &= (1 - \alpha) \int_{\mathbf{R}^n} p \left\| \frac{\nabla p}{p} - \frac{(1 - \alpha)\nabla p + \alpha\nabla q}{(1 - \alpha)p + \alpha q} \right\|^2 dx \\ &+ \alpha \int_{\mathbf{R}^n} q \left\| \frac{\nabla q}{q} - \frac{(1 - \alpha)\nabla p + \alpha\nabla q}{(1 - \alpha)p + \alpha q} \right\|^2 dx \\ &= (1 - \alpha) \int_{\mathbf{R}^n} \frac{\|\nabla p\|^2}{p} dx + \alpha \int_{\mathbf{R}^n} \frac{\|\nabla q\|^2}{q} dx - \int_{\mathbf{R}^n} \frac{\|(1 - \alpha)\nabla p + \alpha\nabla q\|^2}{(1 - \alpha)p + \alpha q} dx \\ &= (1 - \alpha)I(p) + \alpha I(q) - I((1 - \alpha)p + \alpha q), \end{aligned} \quad (16)$$

which is the same form as Equation (10) as desired. From the definitions Equation (13) and Equation (10),  $JFD_{\alpha}(p, q)$  is clearly positive, and vanishes when  $p = q$ . This property also means that Fisher information  $I(p)$  is a convex function. We easily find that this divergence measure does not keep symmetry with respect to the exchange of  $p$  and  $q$ ; that is,  $JFD_{\alpha}(p, q) \neq JFD_{\alpha}(q, p)$ . Instead, under the index change  $\alpha \leftrightarrow 1 - \alpha$ , the following relation is satisfied:

$$JFD_{1-\alpha}(p, q) = JFD_{\alpha}(q, p). \quad (17)$$

Next, we find that an immediate consequence of the application of the Blachman–Stam inequality gives an interesting lower bound for  $JFD_{\alpha}(p, q)$ . The Blachman–Stam inequality [11,41] asserts that, for all  $\alpha \in [0, 1]$ , Fisher information of  $p$  and  $q$  satisfies an inequality

$$\alpha I(p) + (1 - \alpha)I(q) \geq I(\sqrt{\alpha}p + \sqrt{1 - \alpha}q). \quad (18)$$

Therefore, we find the following bound:

**Proposition 2.** The skewed Jensen–Fisher divergence of order  $\alpha$  is lower bounded by the difference of two Fisher information as follows:

$$JFD_{\alpha}(p, q) \geq I(\sqrt{1 - \alpha}p + \sqrt{\alpha}q) - I((1 - \alpha)p + \alpha q). \quad (19)$$

From the positivity of  $JFD_{\alpha}(p, q)$  because of the definition Equation (15) and of the positivity of the relative Fisher information, the above inequality also means an interesting relation  $I(\sqrt{1 - \alpha}p + \sqrt{\alpha}q) \geq I((1 - \alpha)p + \alpha q)$ .

**Remark 3.** A similar bound can be obtained for the skewed Jensen–Shannon divergence of order  $\alpha$  by way of the Shannon–Stam inequality [11]:

$$(1 - \alpha)S(p) + \alpha S(q) \geq S(\sqrt{1 - \alpha}p + \sqrt{\alpha}q).$$

Applying this to the definition Equation (1), we readily obtain an upper bound instead of the lower one

$$JSD_{\alpha}(p, q) \leq S((1 - \alpha)p + \alpha q) - S(\sqrt{1 - \alpha}p + \sqrt{\alpha}q).$$

From the positivity of  $JSD_{\alpha}(p, q)$ , this inequality also means the relation  $S((1 - \alpha)p + \alpha q) \geq S(\sqrt{1 - \alpha}p + \sqrt{\alpha}q)$ .

#### 4. A Bound by Skewed Jensen–Shannon Divergence

It will be useful to relate a divergence measure to another existing one in terms of a bound because the convergence (or similarity) in a divergence can be stronger or weaker than convergence in other measures. This section aims at providing such an example. We consider a lower bound from an integral inequality for gradient functions  $f$ . Contrary to the Sobolev inequality, the logarithmic Sobolev inequality does not depend on dimension  $n$ , in which we consider  $f$  as probability density functions. Recalling that when all functions  $f$  and the distributional gradient  $\nabla f$  belong to  $L^2$ , the probability density function  $\rho(x)$  on  $\mathbf{R}^n$  satisfies the Gross' logarithmic Sobolev inequality [42]

$$\int_{\mathbf{R}^n} |f|^2 \log |f| \rho(x) dx \leq c \int_{\mathbf{R}^n} \|\nabla f\|^2 \rho(x) dx + \|f\|_2^2 \log \|f\|_2^2, \quad (20)$$

where the constant  $c$  is independent of  $f$ . As a special case, when  $\rho(x)$  is the standard Gaussian density with zero mean and unit variance, we have  $c = 1$  (the so-called LSI(1)). Thus, by applying the above by identifying  $f = \sqrt{p/q}$  and  $\rho(x) = q(x)$ , we have an inequality

$$\frac{c}{4} I(p|q) \geq \frac{1}{2} KL(p||q). \quad (21)$$

Therefore, the skewed Jensen–Fisher divergence of order  $\alpha$  is lower bounded by the skewed Jensen–Shannon divergence, as

$$\begin{aligned} JFD_{\alpha}(p, q) &\geq \frac{2}{c} (1 - \alpha) KL(p|| (1 - \alpha)p + \alpha q) + \frac{2}{c} \alpha KL(q|| (1 - \alpha)p + \alpha q) \\ &= \frac{2}{c} JSD_{\alpha}(p, q). \end{aligned} \quad (22)$$

This lower bound indicates that, when we can set the constant as  $c \geq 2$ , we have a tighter distance than the skewed Jensen–Shannon divergence gives.

#### 5. A Lower Bound by Variational Distance

The purpose of this section is to find a bound of the skewed Jensen–Fisher divergence of order  $\alpha$  in terms of the variational distance between  $p$  and  $q$ . We apply the Poincaré inequality to bound  $JFD_{\alpha}(p, q)$  from below. It states that the probability density function  $\rho(x)$  on  $\mathbf{R}^n$  satisfies the inequality (e.g., [31])

$$\int_{\mathbf{R}^n} \left( f(x) - \int_{\mathbf{R}^n} f(y) dy \right)^2 \rho(x) dx \leq \frac{1}{\lambda} \int_{\mathbf{R}^n} \|\nabla f\|^2 \rho(x) dx \quad (23)$$

for all functions  $f$  that belong to  $L^2$ , such that its distributional gradient also belongs to  $\nabla f \in L^2$ . The Poincaré constant or the spectral gap  $\lambda$  is a positive constant.

**Proposition 3.** *The skewed Jensen–Fisher divergence of order  $\alpha$  is lower bounded by the variational distance  $V$  as*

$$JFD_{\alpha}(p, q) \geq \frac{\lambda}{2} \alpha^2 V^2. \quad (24)$$

**Proof.** Identifying  $f = \sqrt{p/q}$  and  $\rho(x) = q$  in the Poincaré inequality, the right-hand side becomes  $I(p|q)/4\lambda$ . On the other hand, we evaluate the left-hand side as

$$1 - \left( \int_{\mathbf{R}^n} \sqrt{pq} dy \right)^2 \geq \frac{1}{2} \int_{\mathbf{R}^n} (\sqrt{p} - \sqrt{q})^2 dy, \quad (25)$$

where we have used the magnitude relationship  $\int \sqrt{pq} dy \geq (\int \sqrt{pq} dy)^2$ . In this sense, the bound becomes loose here, and it states

$$\frac{1}{4\lambda} I(p|q) \geq \frac{1}{2} \int_{\mathbf{R}^n} (\sqrt{p} - \sqrt{q})^2 dy. \quad (26)$$

We want to bound  $JFD_\alpha(p, q)$  in terms of the variational distance  $V$ :

$$V = \int_{\mathbf{R}^n} |p(x) - q(x)| dx. \quad (27)$$

Applying the Schwarz inequality, one has  $V^2 \leq (\int (\sqrt{p} - \sqrt{q})^2 dx) (\int (\sqrt{p} + \sqrt{q})^2 dx)$ . Thus, the second factor of the right-hand side can be bounded by 4 due to  $\int \sqrt{pq} dx \leq 1$ . Hence, the right-hand side of Equation (26), which is the squared Hellinger distance aside from  $1/2$ , is lower bounded by  $V^2/8$ . Regarding  $q$  as  $(1 - \alpha)p + \alpha q$  in this inequality, the skewed relative Fisher information is lower bounded as

$$I(p|(1 - \alpha)p + \alpha q) \geq \frac{\lambda}{2} \alpha^2 V^2. \quad (28)$$

Finally, from the definition of  $JFD_\alpha(p, q)$  (i.e., Equation (15)), we have the lower bound as desired.  $\square$

## 6. Summary and Discussion

We have defined the skewed Jensen–Fisher divergence of order  $\alpha$  and found some bounds. These bounds come from inequalities in which the broadness of a function in terms of  $L^q$  norm bounds an average gradient of a function from below in suitable ways. The logarithmic Sobolev inequality is a tool that provides such a link with the skewed Jensen–Fisher divergence through the relative Fisher information. We know that the Sobolev inequality (not the logarithmic) is also an integral inequality that involves the gradient of a function. Therefore, one may think that it offers a bound for our divergence measure because the Sobolev inequality also relates the  $L^2$  norm of the gradient of a function with an  $L^r$  norm of the function, where  $r$  is an index that depends on the dimension of space  $n$ . However, an unthinking application of it leads to inappropriate bounds. To see this aspect clearly, we consider the case  $n \geq 3$  and see first that  $JFD_\alpha(p, q)$  is lower bounded in terms of the Chernoff  $\alpha$  distance; then, we see that the derived bound loses its effectiveness.

When  $n \geq 3$ , the Sobolev inequality states [43]

$$\|\nabla f\|_2^2 \geq S_n \|f\|_r^2, \quad S_n = \frac{n(n-2)2^{\frac{2}{n}} \pi^{1+\frac{1}{n}}}{4\Gamma((n+1)/2)^{\frac{2}{n}}}, \quad (29)$$

where  $r$  is the Sobolev exponent  $r = 2n/(n-2)$ . Identifying  $f = \sqrt{p/q}$  and taking the measure for the norms as  $q(x)dx$ , we have

$$\|\nabla f\|_2^2 = \frac{1}{4} I(p|q), \quad \|f\|_r^2 = \left( \int_{\mathbf{R}^n} p^{\frac{r}{2}} q^{1-\frac{r}{2}} dx \right)^{2/r}. \quad (30)$$



Note that the power exponent appearing in the latter is  $r/2 = n/(n-2) > 1$ . We recall here that the Chernoff  $\alpha$  distance [44] between two density functions  $p$  and  $q$  is defined as (we here use  $\beta$  instead of  $\alpha$  to avoid confusion with  $\alpha$  for  $JFD_\alpha(p, q)$ )

$$C_\beta(p, q) = -\log\left(\int_{\mathbf{R}^n} p^\beta q^{1-\beta} dx\right). \quad (31)$$

This is also called the skewed Bhattacharyya distance [45], because when  $\beta = 1/2$ , it reduces to the Bhattacharyya distance [46]. The distance  $C_\beta(p, q)$  becomes zero when  $p = q$ , and takes positive values for  $0 \leq \beta \leq 1$ ; however, it can be negative for other ranges, which is our present case. Indeed, for  $\beta \geq 1$ ,  $C_\beta(p, q) \leq 0$  follows, because  $C_\beta(p, q)$  is concave with respect to  $\beta$  and  $C_0(p, q) = C_1(p, q) = 0$ . Thus, rewriting  $\int_{\mathbf{R}^n} p^\beta q^{1-\beta} dx = e^{-C_\beta(p, q)}$ , the Sobolev inequality provides

$$\frac{1}{4}I(p|q) \geq S_n \exp\left[-\frac{n-2}{n}C_{\frac{n}{n-2}}(p, q)\right]. \quad (32)$$

Therefore, identifying this  $q$  as  $\overline{pq} := (1-\alpha)p + \alpha q$  and from the definition of Equation (15), we have the following lower bound in terms of the Chernoff  $\alpha$  distance:

$$JFD_\alpha(p, q) \geq 4S_n \left\{ (1-\alpha) \exp\left[-\frac{n-2}{n}C_{\frac{n}{n-2}}(p, \overline{pq})\right] + \alpha \exp\left[-\frac{n-2}{n}C_{\frac{n}{n-2}}(q, \overline{pq})\right] \right\}, \quad (n \geq 3). \quad (33)$$

The value of the Chernoff distance becomes negatively large as  $p$  and  $q$  differ largely, and when we consider a higher dimensional space. Indeed, the coefficient  $S_n$  is an increasing function of  $n$ , which starts from  $S_3 = 5.4779\dots$ . Even when the Chernoff distance vanishes by increasing  $n$  (recall  $C_1(p, q) = 0$ ) or by approaching the one density function  $q$  to the other one  $p$ , Equation (33) asserts the finitely large lower bound  $4S_n$ . Thus, the derived lower bound loses its effectiveness. The same argument is true for two- and one-dimensional cases. Where does this inexpedience come from? The Sobolev inequality holds for functions  $f$  that have compact support (thus vanishes at infinity) with  $\nabla f \in L^2(\mathbf{R}^n)$ . On the other hand, the log-Sobolev inequality holds for a class of functions with  $f \in L^2(\mathbf{R}^n)$  and  $\nabla f \in L^2(\mathbf{R}^n)$ . The density functions vanish at infinity. However, when we choose  $f = \sqrt{p/q}$  to make  $I(p|q)$  (and  $\sqrt{q/p}$  for  $I(q|p)$ ), the ratio  $p/q$  (and  $q/p$ ) does not vanish when  $|x| \rightarrow \infty$  or at the boundary. This fact indicates that the Sobolev inequality cannot be applied in our present study.

**Funding:** The author did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

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

**Informed Consent Statement:** Not applicable.

**Acknowledgments:** The author appreciates two anonymous reviewers for useful comments for the revision of the manuscript. The author thanks Managing Editor Sonic Zhao for waiving the APC for publication.

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

## References

1. Lin, J. Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory* **1991**, *37*, 145. [CrossRef]
2. Nielsen, F. A family of statistical symmetric divergences based on Jensen's inequality. *arXiv* **2011**, arXiv:1009.4004v2.
3. Yamano, T. Some bounds for skewed  $\alpha$ -Jensen-Shannon divergence. *Results Appl. Math.* **2019**, *3*, 10064. [CrossRef]
4. Sánchez-Moreno, P.; Zarzo, A.; Dehesa, J.S. Jensen divergence based on Fisher's information, *J. Phys. A Math. Theor.* **2012**, *45*, 125305. [CrossRef]
5. Lee, L. Measures of distributional similarity. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics, University of Maryland, College Park, MA, USA, 20–26 June 1999; pp. 25–32.



6. Lee, L. On the effectiveness of the skew divergence for statistical language analysis. In *Artificial Intelligence and Statistics*; Morgan Kaufmann Publisher: Burlington, MA, USA, 2001; pp. 65–72.
7. Sibson, R. Information radius. *Z. Wahrscheinlichkeitstheorie Verw Geb.* **1969**, *14*, 149. [[CrossRef](#)]
8. Endres, D.; Schindelin, J. A new metric for probability distributions. *IEEE Trans. Inf. Theory* **2003**, *49*, 1858. [[CrossRef](#)]
9. Lin, J.; Wong, S.K.M. A new directed divergence measure and its characterization. *Int. J. Gen. Syst.* **1990**, *17*, 73. [[CrossRef](#)]
10. Nielsen, F.; Nock, R. On the geometry of mixtures of prescribed distributions. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 2861–2865.
11. Stam, A. Some Inequalities Satisfied by the Quantities of Information of Fisher and Shannon. *Inf. Control* **1959**, *2*, 101. [[CrossRef](#)]
12. Dembo, A.; Cover, T.; Thomas, J. Information theoretic inequalities. *IEEE Trans. Inf. Theory* **1991**, *37*, 1501. [[CrossRef](#)]
13. Cover, T.; Thomas, J. *Elements of Information Theory*, 2nd ed.; Wiley-Interscience: New York, NY, USA, 2006.
14. Narayanan, K.R.; Srinivasa, A.R. On the thermodynamic temperature of a general distribution. *arXiv* **2007**, arXiv:0711.1460v2.
15. Fisher, R.A. Theory of statistical estimation. *Proc. Camb. Philos. Soc.* **1925**, *22*, 700. [[CrossRef](#)]
16. Rao, C.R. *Linear Statistical Inference and Its Applications*; Wiley: New York, NY, USA, 1965.
17. Yamano, T. Relative Fisher information of hydrogen-like atoms. *Chem. Phys. Lett.* **2018**, *691*, 196. [[CrossRef](#)]
18. Yamano, T. Fisher information of radial wavefunctions for relativistic hydrogenic atoms. *Chem. Phys. Lett.* **2019**, *731*, 136618. [[CrossRef](#)]
19. Yamano, T. de Bruijn-type identity for systems with flux. *Eur. Phys. J. B* **2013**, *86*, 363. [[CrossRef](#)]
20. Hyvärinen, A. Estimation of non-normalized statistical models by score matching. *J. Mach. Learn. Res.* **2005**, *6*, 695.
21. Hyvärinen, A. Some extensions of score matching. *Comput. Stat. Data Anal.* **2007**, *51*, 2499. [[CrossRef](#)]
22. Yang, Y.; Martin, R.; Bondell, H. Variational approximations using Fisher divergence. *arXiv* **2019**, arXiv:1905.05284v1.
23. Huggins, J. H.; Campbell, T.; Kasprzak, M.; Broderick, T. Practical bounds on the error of Bayesian posterior approximations: A nonasymptotic approach. *arXiv* **2018**, arXiv:1809.09505.
24. Elkhailil, K.; Hasan, A.; Ding, J.; Farsiu, S.; Tarokh, V. Fisher Auto-Encoders, *Proc. Mach. Learn. Res.* **2021**, *130*, 352.
25. Kostrikov, I.; Tompson, J.; Fergus, R.; Nachum, O. Offline reinforcement learning with Fisher divergence critic regularization. *Proc. Mach. Learn. Res.* **2021**, *139*, 5774.
26. Hammad, P. Mesure d'ordre  $\alpha$  de l'information au sens de Fisher. *Rev. Stat. Appl.* **1978**, *26*, 73.
27. Barron, A.R. Entropy and the central limit theorem. *Ann. Prob.* **1986**, *14*, 336. [[CrossRef](#)]
28. Johnson, O.; Barron, A. Fisher information inequalities and the central limit theorem. *Probab. Theory Relat. Fields* **2004**, *129*, 391. [[CrossRef](#)]
29. Johnson, O.T. *Information Theory and the Central Limit Theorem*; World Scientific: Singapore, 2004; pp. 23–24.
30. Otto, F.; Villani, C. Generalization of an Inequality by Talagrand and Links with the Logarithmic Sobolev Inequality. *J. Funct. Anal.* **2000**, *173*, 361. [[CrossRef](#)]
31. Villani, C. *Topics in Optimal Transportation, Graduate Studies in Mathematics*; American Mathematical Society: Providence, RI, USA, 2000; Volume 58, p. 278.
32. Antolín, J.; Angulo, J.C.; López-Rosa, S. Fisher and Jensen-Shannon divergences: Quantitative comparisons among distributions. Application to position and momentum atomic densities. *J. Chem. Phys.* **2009**, *130*, 074110. [[CrossRef](#)]
33. López-Rosa, S.; Antolín, J.; Angulo, J.C.; Esquivel, R.O. Divergence analysis of atomic ionization processes and isoelectronic series. *Phys. Rev. A* **2009**, *80*, 012505. [[CrossRef](#)]
34. Mukherjee, N.; Roy, A.K. Relative Fisher information in some central potentials. *Ann. Phys.* **2018**, *398*, 190. [[CrossRef](#)]
35. Yamano, T. Relative Fisher information for Morse potential and isotropic quantum oscillators. *J. Phys. Commun.* **2018**, *2*, 085018. [[CrossRef](#)]
36. Yamano, T. Fisher Information of Free-Electron Landau States. *Entropy* **2021**, *23*, 268. [[CrossRef](#)] [[PubMed](#)]
37. Levämäki, H.; Nagy, Á.; Vilja, I.; Kokko, K.; Vitos, L. Kullback-Leibler and relative Fisher information as descriptors of locality. *Int. J. Quantum Chem.* **2018**, *118*, e25557; doi: 10.1002/qua.2557. [[CrossRef](#)]
38. Nagy, Á. Relative information in excited-state orbital-free density functional theory. *Int. J. Quantum Chem.* **2020**, *120*, e26405. [[CrossRef](#)]
39. Yamano, T. Phase space gradient of dissipated work and information: A role of relative Fisher information. *J. Math. Phys.* **2013**, *54*, 113301. [[CrossRef](#)]
40. Yamano, T. Constraints on stochastic heat probability prescribed by exchange fluctuation theorems. *Results Phys.* **2020**, *18*, 103300. [[CrossRef](#)]
41. Blachman, K.M. The convolution inequality for entropy powers. *IEEE Trans. Inf. Theory* **1965**, *2*, 267. [[CrossRef](#)]
42. Gross, L. Logarithmic Sobolev inequalities. *Am. J. Math.* **1975**, *97*, 1061. [[CrossRef](#)]
43. Lieb, E.H.; Loss, M. *Analysis*, 2nd ed.; Graduate Studies in Mathematics; American Mathematical Society: Providence, RI, USA, 2001; Chapter 8, Volume 14.
44. Chernoff, H. A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on the sum of Observations. *Ann. Math. Stat.* **1952**, *23*, 493. [[CrossRef](#)]
45. Nielsen, F.; Boltz, S. The Burbea-Rao and Bhattacharyya centroids. *IEEE Trans. Inf. Theory* **2011**, *57*, 5455. [[CrossRef](#)]
46. Bhattacharyya, A. On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math. Soc.* **1943**, *35*, 99.