Entropy **2014**, *16*, 2611-2628; doi:10.3390/e16052611

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Article

Scale-Invariant Divergences for Density Functions

Takafumi Kanamori

Nagoya University, Furocho, Chikusaku, Nagoya 464-8603, Japan;
E-Mail: kanamori@is.nagoya-u.ac.jp; Tel.: +81-52-789-4598.

Received: 6 January 2014; in revised form: 26 April 2014 / Accepted: 28 April 2014 / Published: 13 May 2014

Abstract: Divergence is a discrepancy measure between two objects, such as functions, vectors, matrices, and so forth. In particular, divergences defined on probability distributions are widely employed in probabilistic forecasting. As the dissimilarity measure, the divergence should satisfy some conditions. In this paper, we consider two conditions: The first one is the scale-invariance property and the second is that the divergence is approximated by the sample mean of a loss function. The first requirement is an important feature for dissimilarity measures. The divergence will depend on which system of measurements we used to measure the objects. Scale-invariant divergence is transformed in a consistent way when the system of measurements is changed to the other one. The second requirement is formalized such that the divergence is expressed by using the so-called composite score. We study the relation between composite scores and scale-invariant divergences, and we propose a new class of divergences called Hölder divergence that satisfies two conditions above. We present some theoretical properties of Hölder divergence. We show that Hölder divergence unifies existing divergences from the viewpoint of scale-invariance.

Keywords: divergence; scale invariance; composite score; Hölder inequality; reverse Hölder inequality

---

1. Introduction

Nowadays, divergence measures are ubiquitous in the field of information sciences. The divergence is a discrepancy measure between two objects, such as functions, vectors, matrices, and so forth. In particular, divergences defined on the set of probability distributions are widely used for probabilistic forecasting such as weather and climate prediction [1,2], computational fiance [3], and so forth. In many statistical inferences, statistical models are prepared to estimate the probability distribution generating

observed samples. A divergence measure between the true probability and the statistical model is estimated based on observed samples, and the probability distribution in the statistical model that minimizes the divergence measure is chosen as the estimator. A typical example is the maximum likelihood estimator based on the Kullback-Leibler divergence [4].

Dissimilarity measures for statistical inference should satisfy some conditions. In this paper, we focus on two conditions. The first one is the scale-invariance property, and the second one is that the divergence should be represented by using the so-called composite score [5], that is an extension of scores [6].

The first requirement is the scale-invariance. Suppose that the divergence is used to measure the dissimilarity between two objects, then the divergence will depend on the system of measurements we used to measure the objects. The scale-invariant divergence has a favorable property such that it is transformed in a consistent way when the system of measurements is changed to the other one. For example, the measured value between two objects depends on the unit of length. Typically, measured values in different units are transformed to each other by multiplying an appropriate positive constant. The Kullback-Leibler divergence that is one of the most popular divergences has the scale-invariance property for the measurement of training samples [7].

As the second requirement, dissimilarity measures should be expressed as the form of composite scores. This is a useful property, when the divergence is employed for the statistical inference of the probability densities. When the divergence $D(f, g)$ is calculated through the expectation with respect to the probability density f, the sample mean over the observations works to approximate the divergence. The score [2,5,6,8–10] is the class of dissimilarity measures that are calculated through the sample mean of the observed data. The characterization of the score is studied by [6,10], and the deep connection between scores and divergences were revealed.

In the present paper, we propose composite scores as an extension of scores, and study the relation between composite scores and scale-invariant divergences. We propose a new class of divergences called Hölder divergence, that is defined through a class of composite scores. We show that Hölder divergence unifies existing divergences from the viewpoint of the scale-invariance. The Hölder divergence with the one-dimension parameter γ is defined from a function ϕ. Partially, the Hölder divergence with non-negative γ was proposed in [5]. Here, we extend the previous result to any real number γ.

The remainder of the article is as follows: In Section 2, some basic notions such as divergence, scale-invariance and score are introduced. In Section 3, we propose the Hölder divergence. Some theoretical properties of Hölder divergence are investigated in Section 4. In Section 5, we close this article with a discussion of the possibility of the newly introduced divergences. Technical calculations and proofs are found in the appendix.

Let us summarize the notations to be used throughout the paper. Let $\mathbb{R}$ be the set of all real numbers, $\mathbb{R}_+$ be the set of all non-negative real numbers, and $\mathbb{R}_{++}$, and the set of all positive real numbers. For a real-valued function $f : \Omega \to \mathbb{R}$ defined on a domain Ω in the Euclidean space, let $\langle f \rangle$ be the integral $\int_\Omega f(x)dx$. In most arguments of the current paper, Ω is the closed interval $[0, 1]$ in $\mathbb{R}$. Extension of the theoretical results to any compact set in the multi-dimensional Euclidean space is straightforward.

2. Preliminaries

In this section, we show definitions of some basic concepts.

2.1. Divergences and Scores

Let us introduce scores and divergences. Below, positive-valued functions are defined on the compact set Ω. The *score* is defined as the real-valued functional $S(f,g)$ in which $f(x)$ and $g(x)$ are positive-valued functions on Ω.

Let $D(f,g)$ be

$$D(f,g) = S(f,g) - S(f,f).$$

The functional $D(f,g)$ is called the *divergence*, if $D(f,g) \geq 0$ holds with equality if and only if $f = g$. Suppose that the score $S(f,g)$ induces a divergence. Then, clearly the score should satisfy $S(f,g) \geq S(f,f)$ with equality only when $f = g$. The divergence does not necessarily satisfy the definition of the distance, because neither the symmetry nor triangle inequality holds in general.

Bregman divergence [11] and Csiszár φ-divergence [12,13] are important classes of divergences. Here we focus on the Bregman divergence, since they are frequently employed in various statistical inferences. See [6,11] for details.

Definition 1 (Bregman divergence; Bregman score)**.** *For positive-valued function $f : \Omega \to \mathbb{R}_{++}$, let $G(f)$ be a strictly convex functional and $G^*_f(x)$ be the functional derivative of G at f, i.e., $G^*_f(x)$ is determined from the equality*

$$\frac{d}{d\varepsilon} G(f + \varepsilon h)\big|_{\varepsilon=0} = \int_\Omega G^*_f(x)h(x)dx = \langle G^*_f h \rangle$$

for any h such that $f + \varepsilon h$ is a positive-valued function for sufficiently small ε. Then, the Bregman divergence is defined as

$$D(f,g) = G(f) - G(g) - \langle G^*_g(f-g) \rangle.$$

The score associated with the Bregman divergence is called the Bregman score, that is defined as

$$S(f,g) = -G(g) - \langle G^*_g(f-g) \rangle.$$

The functional G is referred to as the potential of the Bregman divergence, and it satisfies the equality $G(f) = -S(f,f)$.

The rigorous definition of G^*_f requires the dual space of Banach space. See [14] (Chapter 4) for sufficient conditions of the existence of G^*_f. To avoid technical difficulties, we assume the existence of the functional derivative in the above definition.

The remarkable property of Bregman divergence is that associated score $S(f,g)$ is represented as the linear function of f. This is a nice property for statistical inference, since one can substitute the empirical distribution directly into f. In other words, the sample-based approximation of the Bregman score is obtained by the sample-mean of a function depending on the model g. For this reason, the

Bregman divergences have a wide range of applications in statistics, machine learning, data mining, and so forth [15–17].

Though Bregman divergence is a popular class of divergences, the computation of the potential may be a hard task. The separable Bregman divergence is an important subclass of Bregman divergences. In many applications of statistical forecasting, the separable Bregman divergences are used due to the computational tractability.

Definition 2 (separable Bregman divergence)**.** *Let $J : \mathbb{R}_+ \to \mathbb{R}$ be a strictly convex differentiable function. The separable Bregman score is defined as*

$$S(f,g) = -\int_\Omega \{J(g(x)) + J'(g(x))(f(x) - g(x))\}\, dx = -\langle J(g) + J'(g)\,(f-g)\rangle,$$

where J' is the derivative of J. The separable Bregman divergence is

$$D(f,g) = S(f,g) - S(f,f) = \langle J(f) - J(g) - J'(g)(f-g)\rangle.$$

The potential of the separable Bregman divergence is $G(f) = \langle J(f)\rangle$.

Due to the convexity of J, the non-negativity of the separable Bregman divergence is guaranteed. Moreover, the strict convexity of J ensures that the equality $D(f,g) = 0$ holds only if $f = g$. Some examples of divergences are shown below.

Example 1 (Kullback-Leibler divergence)**.** *One of the most popular divergences in information sciences is the Kullback-Leibler(KL) divergence [4]. Let us define the KL score for positive-valued functions f and g as*

$$S_{\mathrm{KL}}(f,g) = \langle -f\log g + g\rangle.$$

The associated divergence is called the KL divergence, that is defined as

$$D_{\mathrm{KL}}(f,g) = S_{\mathrm{KL}}(f,g) - S_{\mathrm{KL}}(f,f) = \left\langle f\log\frac{f}{g} - f + g\right\rangle.$$

This is represented as the separable Bregman divergence with the potential $G(f) = \langle f\log f - f\rangle$ defined from $J(z) = z\log z - z$.

Example 2 (Itakura-Saito distance)**.** *The Itakura-Saito (IS) distance was originally used to measure the dissimilarity between two power spectrum densities [18]. Though IS distance does not satisfy the mathematical condition of the distance, the term "distance" is conventionally used. For positive-valued functions f, g on Ω, IS score is defined as*

$$S_{\mathrm{IS}}(f,g) = \left\langle \frac{f}{g} + \log g\right\rangle,$$

and the IS distance is defined as

$$D_{\mathrm{IS}}(f,g) = S_{\mathrm{IS}}(f,g) - S_{\mathrm{IS}}(f,f) = \left\langle \frac{f}{g} - \log\frac{f}{g} - 1\right\rangle.$$

The non-negativity of $D_{\mathrm{IS}}(f,g)$ is guaranteed by the inequality of $z-\log z-1\geq 0$ for $z>0$. The IS distance is the separable Bregman divergence with the potential $G(f)=-\langle\log f+1\rangle$. The IS distance is scale-invariant, i.e., $D_{\mathrm{IS}}(af,ag)=D_{\mathrm{IS}}(f,g)$ holds for any positive real number a. This invariance ensures that the low energy components have the same relative importance as high energy ones. This is especially important in short-term audio spectra [19,20].

Example 3 (density-power divergence)**.** *The density-power divergence is a one-parameter extension of the KL-divergence. The density-power score is defined as*

$$S^{(\gamma)}_{\mathrm{power}}(f,g)=\left\langle\frac{1}{1+\gamma}g^{1+\gamma}-\frac{1}{\gamma}fg^{\gamma}\right\rangle$$

for $\gamma\in\mathbb{R}\setminus\{0,-1\}$, and the density power divergence is defined as

$$D^{(\gamma)}_{\mathrm{power}}(f,g)=S^{(\gamma)}_{\mathrm{power}}(f,g)-S^{(\gamma)}_{\mathrm{power}}(f,f)=\left\langle\frac{1}{1+\gamma}g^{1+\gamma}-\frac{1}{\gamma}fg^{\gamma}+\frac{1}{\gamma(1+\gamma)}f^{1+\gamma}\right\rangle.$$

The density-power divergence is employed in the robust parameter estimation [21,22]. The limit $\gamma\to 0$ of the density-power divergence yields the KL-divergence, and the limit $\gamma\to -1$ yields the IS-distance. Though originally the density-power divergence is defined for positive γ [21], the above definition works for any real number γ. The density-power divergence is expressed as the separable Bregman divergence with the potential $G(f)=\frac{1}{\gamma(1+\gamma)}\langle g^{1+\gamma}\rangle$.

Example 4 (pseudo-spherical divergence; γ divergence)**.** *The pseudo-spherical divergence [6,23] is defined as*

$$D^{(\gamma)}_{\mathrm{sphere}}(f,g)=\frac{1}{\gamma}\langle f^{1+\gamma}\rangle^{1/(1+\gamma)}-\frac{\langle fg^{\gamma}\rangle}{\gamma\langle g^{1+\gamma}\rangle^{\gamma/(1+\gamma)}},\quad \gamma\neq 0,-1,$$

that is derived from the pseudo-spherical score

$$S^{(\gamma)}_{\mathrm{sphere}}(f,g)=-\frac{\langle fg^{\gamma}\rangle}{\gamma\langle g^{1+\gamma}\rangle^{\gamma/(1+\gamma)}}.$$

The pseudo-spherical divergence does note satisfy the definition of the divergence in the present paper, since $D^{(\gamma)}_{\mathrm{sphere}}(f,g)=0$ holds for linearly dependent functions f and g. On the set of probability density functions, however, the equality $D^{(\gamma)}_{\mathrm{sphere}}(p,q)=0$ leads to $p=q$. Thus, pseudo-spherical divergence is still useful in statistical inference, though it is not divergence on the set of positive-valued functions. The γ divergence [24] is defined as $-\log(-S^{(\gamma)}_{\mathrm{sphere}}(f,g))+\log(-S^{(\gamma)}_{\mathrm{sphere}}(f,f))$, and the first term of the γ divergence is used for robust parameter estimation. The pseudo-spherical divergence is represented as the non-separable Bregman divergence with the potential $G(f)=\frac{1}{\gamma}\langle f^{1+\gamma}\rangle^{1/(1+\gamma)}$. This potential is strictly convex on the set of probability densities. The parameter γ can take both positive and negative real numbers.

Example 5 (α-divergence)**.** *For positive-valued functions f,g, the α-divergence [25,26] is defined as*

$$D^{(\alpha)}_{\mathrm{alpha}}(f,g)=\frac{1}{\alpha(\alpha-1)}\langle f^{\alpha}g^{1-\alpha}-\alpha f-(1-\alpha)g\rangle$$

for $\alpha \in \mathbb{R}\backslash\{0,1\}$. Generally, α-divergence is not included in Bregman divergence, since the term $f^\alpha g^{1-\alpha}$ is not linear in f. The limit $\alpha \to 1$ and $\alpha \to 0$ yield the KL-divergence $D_{\mathrm{KL}}(f,g)$ and $D_{\mathrm{KL}}(g,f)$, respectively. We show that the α-divergence has the invariance property. Let $c(x)$ be a one-to-one differentiable mapping from $x \in \mathbb{R}^d$ to $c(x) \in \mathbb{R}^d$, and $c'(x) \in \mathbb{R}^d$ be the gradient vector. For the transformation $f(x) \mapsto f_c(x) = |c'(x)|f(c(x))$, the equality $D^{(\alpha)}_{\mathrm{alpha}}(f_c, g_c) = D^{(\alpha)}_{\mathrm{alpha}}(f,g)$ holds. This invariance is a common property of Csiszár's φ-divergence [12,13].

2.2. Scale-Invariance of Divergences

Let us consider the scale-invariance of divergences. Suppose that $f(x)$ is a density at the point $x \in \Omega = [0,1]$. Here, not only the probability density but also the mass density or spectrum density is considered. Hence, the density is not necessarily normalized, but should be finite measures. For density functions, the total mass does not change under the variable transformation of the coordinate x. Especially, for the scale-transformation $x \mapsto y = x/\sigma$ with $\sigma > 0$, the density $f(x)$ in the x-coordinate should be transformed to $\sigma f(\sigma y)$ in the y-coordinate. In addition, under the scale-transformation of the function value, the density $f(x)$ is transformed to $af(x)$ with some positive constant a. For the density function, we allow the combination of the above two transformations,

$$f(x) \longmapsto f_{a,\sigma}(x) = a\sigma f(\sigma x), \quad a, \sigma > 0. \tag{1}$$

The support of f is also properly transformed to that of $f_{a,\sigma}$. The transformation Equation (1) is induced by changing the unit of systems of the measurement. On multi-dimensional space, the density $f_{a,\sigma}(x)$ with the positive constant a and invertible matrix σ is defined as $a|\det\sigma|f(\sigma x)$, in which $\det\sigma$ is the determinant of the matrix σ. In most arguments in the paper, one-dimensional case is considered, since the extension to the multi-dimensional domain is straightforward.

As a natural requirement, the divergence measure should not be essentially affected by systems of measurement. More concretely, the relative nearness between two densities should be preserved under the transformation Equation (1). This requirement is formalized as the relative invariance for the scale transformation, *i.e.*, there exists a function $\kappa(a,\sigma)$ such that the equality

$$D(f_{a,\sigma}, g_{a,\sigma}) = \kappa(a,\sigma)D(f,g) \tag{2}$$

holds for any pair of densities f, g and any transformation $f \mapsto f_{a,\sigma}$. The divergence satisfying Equation (2) is referred to as the scale-invariant divergence. Some popular divergences satisfy the scale-invariance; $\kappa(a,\sigma) = a$ for KL-divergence and α-divergence, $\kappa(a,\sigma) = \sigma^{-1}$ for IS-distance, $\kappa(a,\sigma) = a^{1+\gamma}\sigma^\gamma$ for density-power divergence, and $\kappa(a,\sigma) = a\sigma^{\gamma/(1+\gamma)}$ for pseudo-spherical divergence.

2.3. Divergence for Statistical Inference

The divergence $D(f,g)$ or score $S(f,g)$ is widely applied in statistical inference. The discrepancy between two probability densities are measured by the divergence or score. Typically, the true probability density p and the model probability density q are substituted into the divergence $D(p,q)$, and $D(p,q)$ is minimized with respect to the model q in order to estimate the probability density p. This is the same as

the minimization of the score $S(p,q)$. Usually, one cannot directly access the true probability density. However, the true probability p can be replaced with the empirical probability density of observed samples, when the samples are observed from p. Given the empirical probability density $\widetilde{p}$, the empirical score $S(\widetilde{p},q)$ is expected to approximate $S(p,q)$. The estimator is obtained by minimizing $S(\widetilde{p},q)$ with respect to the model density q.

Generally, one cannot directly substitute the empirical probability density $\widetilde{p}$ into p of the score $S(p,q)$, since $\widetilde{p}$ is expressed by the sum of Dirac's delta function. Suppose that the score depends on p through the expectation of a random variable with respect to p. Then, one can substitute the empirical distribution $\widetilde{p}$ into p. Let us introduce the composite score into which one can substitute the empirical distributions.

Definition 3 (composite score)**.** *For positive-valued functions f and g on Ω, the score expressed as*

$$S(f,g)=\psi(\langle fU(g)\rangle,\,\langle V(g)\rangle)$$

is called the composite score, where ψ is a real-valued function on $\mathbb{R}^2$ and U and V are real-valued functions. The integrals $\langle fU(g)\rangle$ and $\langle V(g)\rangle$ denote $\int_\Omega f(x)U(g(x))dx$ and $\int_\Omega V(g(x))dx$, respectively.

The composite score was introduced in [5]. The function ψ is arbitrary in the above definition. When we impose some constraints on the composite scores, the form of ψ will be restricted. Concrete expressions of ψ are presented in Section 3. For the purpose of statistical inference, it is sufficient to define scores on the set of probability densities. However, the scores defined for positive-valued functions are useful to investigate theoretical properties; see [10] for details.

Separable Bregman divergences are represented by using composite scores. Indeed, the separable Bregman divergence with the potential $G(g)=\langle J(g)\rangle$ is obtained by setting $U(g)=-J'(g)$, $V(g)=J'(g)g-J(g)$ and $\psi(a,b)=a+b$ in the composite score. Hence, the KL-divergence, Itakura-Saito distance, density-power divergence are represented by using the composite score. Though the pseudo-spherical divergence over the set of probability densities is a non-separable Bregman divergence, it is expressed by the composite score as shown in Section 3.

Scale-invariant divergences defined from composite scores are useful for statistical inference. Suppose that $D(f,g)=S(f,g)-S(f,f)$ is the scale-invariant divergence. Then, the statistical inference using the score $S(f,g)$ does not essentially depend on the systems of measurement in the observations. Let $\widehat{q}$ be the estimator based on the sample x, and $\widehat{(q_{1,\sigma})}$ be the estimator based on the transformed sample σx with the model $q_{1,\sigma}$, where $q_{1,\sigma}(x)=\sigma q(\sigma x)$. If the estimator is obtained as the optimal solution of the score that induces the scale-invariant divergence, we obtain $\widehat{(q_{1,\sigma})}=(\widehat{q})_{1,\sigma}$. Such estimator is called the *equivariant* estimator [27]. The estimation result based on the equivariant estimator is transformed in the consistent way, when the systems of the measurement is changed.

Let us define the equivalence class among scores. The two scores are said to be *equivalent* if a score is transformed to the other score by a strictly increasing function, i.e., for any monotone increasing function ξ, two scores, $S(f,g)$ and $\xi(S(f,g))$, are equivalent. The statistical inference is often conducted by minimizing the score. Hence, the equivalent scores provide the same estimator. If a equivalence class includes a score that leads to a scale-invariant divergence, all scores in the class provide the equivariant estimator.

In sequel sections, we introduce the Hölder score that is a class of composite scores with the scale-invariance property. Then, we investigate theoretical properties of the Hölder score.

3. Hölder Divergences

Let us define a class of scale-invariant divergences expressed by the composite score. The divergence is called the Hölder divergence. The name comes from the fact that the Hölder inequality or its reverse variant is used to prove the non-negativity of the divergence. The Hölder divergence unifies existing divergences from the viewpoint of the scale-invariance.

Definition 4 (Hölder score; Hölder divergence)**.** *The Hölder score is defined from a real number $\gamma \in \mathbb{R}$ and a function $\phi : \mathbb{R}_+ \to \mathbb{R}$ as follows:*

- *For $\gamma \neq 0, -1$ and $s \in \{1, -1\}$, the Hölder score is defined as*

$$S_\gamma(f, g) = \phi\left(\frac{\langle fg^\gamma \rangle}{\langle g^{1+\gamma} \rangle}\right) \langle g^{1+\gamma} \rangle^s. \tag{3}$$

  The parameter s is set to $s = 1$ for $\gamma > 0$ or $\gamma < -1$, and $s = -1$ for $0 > \gamma > -1$. The function ϕ satisfies $\phi(1) = -1$ and $\phi(z) \geq -z^{s(1+\gamma)}$ for $z \geq 0$, and the equality holds only when $z = 1$.

- *For $\gamma = 0$, the Hölder score is defined as $S_0(f, g) = \langle -f \log g + g + cf \rangle$, where c is a real number.*

- *For $\gamma = -1$, the Hölder score is defined as $S_{-1}(f, g) = \langle f/g + \log g + cf \rangle$, where c is a real number.*

The Hölder divergence is defined as

$$D_\gamma(f, g) = S_\gamma(f, g) - S_\gamma(f, f)$$

for $\gamma \in \mathbb{R}$.

The Hölder divergence with the non-negative γ is defined in [5]. For $\gamma < 0, \gamma \neq -1$, it is sufficient to define the function $\phi(z)$ for $z > 0$, since the computation of $\phi(0)$ does not required for such γ under the condition that the integral in the divergence is finite. The characterization of the Hölder score is shown in Section 4.3. The Hölder score (divergence) defined from the parameters γ and the function ϕ is denoted as S_γ (D_γ) with ϕ, or S_γ^ϕ (D_γ^ϕ). It is clear that Hölder score is a composite score. We show that Hölder divergence satisfies the conditions of the divergence.

Theorem 1. *For positive-valued functions f, g, the Hölder divergence $D_\gamma(f, g)$ satisfies the inequality $D_\gamma(f, g) \geq 0$ with equality if and only if $f = g$.*

Proof. The Hölder divergences D_0 and D_{-1} coincide with the KL-divergence and IS distance, respectively. Hence, D_γ with $\gamma = 0$ or -1 is the divergence.

For positive-valued functions f and g, the Hölder inequality $\langle fg \rangle \leq \langle f^\alpha \rangle^{1/\alpha} \langle g^\beta \rangle^{1/\beta}$ holds for $1/\alpha + 1/\beta = 1$ with $\alpha, \beta > 1$, and the reverse Hölder inequality $\langle fg \rangle \geq \langle f^\alpha \rangle^{1/\alpha} \langle g^\beta \rangle^{1/\beta}$ holds for $1/\alpha + 1/\beta = 1$ with $1 > \alpha > 0 > \beta$.

For $\gamma \notin [-1,0]$, the Hölder inequality or its reverse variant leads to $\langle fg^\gamma\rangle^{1+\gamma} \leq \langle f^{1+\gamma}\rangle\langle g^{1+\gamma}\rangle^\gamma$. Hence, we have

$$S_\gamma(f,g) \geq -\left(\frac{\langle fg^\gamma\rangle}{\langle g^{1+\gamma}\rangle}\right)^{(1+\gamma)}\langle g^{1+\gamma}\rangle \geq -\langle f^{1+\gamma}\rangle = S_\gamma(f,f),$$

in which the first inequality comes from $\phi(z) \geq -z^{1+\gamma}$ and the second inequality is derived from the (reverse) Hölder inequality. When the first and second inequalities become equality, we have $\langle fg^\gamma\rangle/\langle g^{1+\gamma}\rangle = 1$ and the linearly dependence of f and g. As a result, the equality $S_\gamma(f,g) = S_\gamma(f,f)$ gives $f = g$.

For $\gamma \in (-1,0)$, the reverse Hölder inequality for positive-valued functions f and g is expressed as $\langle fg^\gamma\rangle^{1+\gamma} \geq \langle f^{1+\gamma}\rangle\langle g^{1+\gamma}\rangle^\gamma$. Hence, we have

$$S_\gamma(f,g) \geq -\left(\frac{\langle g^{1+\gamma}\rangle}{\langle fg^\gamma\rangle}\right)^{1+\gamma}\frac{1}{\langle g^{1+\gamma}\rangle} \geq -\frac{1}{\langle f^{1+\gamma}\rangle} = S_\gamma(f,f)$$

in which the first inequality comes from $\phi(z) \geq -z^{-(1+\gamma)}$ and the second inequality is derived from the reverse Hölder inequality. The same argument in the case of $\gamma \notin [-1,0]$ works to show that the equality $S_\gamma(f,g) = S_\gamma(f,f)$ leads to $f = g$. □

The Hölder divergences have the scale-invariance. The following calculation is straightforward.

Theorem 2. *For the Hölder divergence, the equality*

$$D_\gamma(f_{a,\sigma}, g_{a,\sigma}) = (a^{1+\gamma}\sigma^\gamma)^s D_\gamma(f,g), \quad \gamma \in \mathbb{R}\setminus\{0,-1\}$$

holds for $a, \sigma > 0$.

In addition, we have $D_0(f_{a,\sigma}, g_{a,\sigma}) = aD_0(f,g)$ and $D_{-1}(f_{a,\sigma}, g_{a,\sigma}) = \sigma^{-1}D_{-1}(f,g)$. There is no Hölder divergence such that the equality $D_\gamma(f_{a,\sigma}, g_{a,\sigma}) = D_\gamma(f,g)$ holds for arbitrary $a, \alpha > 0$. Moreover, the theorem in Section 4.3 ensures that there is no scale-invariant divergence based on the composite score such that the scale function, $\kappa(a,\sigma)$, is constant.

The class of Hölder divergences includes some popular divergences that are used in statistics and information theory. Some examples are shown below.

Example 6 (density-power divergence and Hölder divergence)**.** *The Hölder score with* $\gamma \notin [-1,0]$ *and* $\phi(z) = -(1+\gamma)z + \gamma$ *is equivalent with the density-power score in Example 3 with the same* γ. *For* $\gamma \in (-1,0)$, *the Hölder score with* $\phi(z) = -1/((1+\gamma)z - \gamma)$ *is equivalent with the density-power score with the same* γ.

Example 7 (pseudo-spherical divergence and Hölder divergence)**.** *The pseudo-spherical score is equivalent with the score Equation (3) with* $\phi(z) = -z^{s(1+\gamma)}$, *where* $s = 1$ *for* $\gamma \notin [-1,0]$ *and* $s = -1$ *for* $\gamma \in (-1,0)$. *In our definition of the Hölder score, setting* $\phi(z) = -z^{s(1+\gamma)}$ *is not allowed.*

Example 8 (Bregman-Hölder divergence)**.** *For* $\gamma \neq 0, -1$ *and* $\kappa \neq 0, 1$, *let us define the potential* $G_{\gamma,\kappa}(f)$ *as*

$$G_{\gamma,\kappa}(f) = \begin{cases} \langle f^{1+\gamma}\rangle^{\kappa/(1+\gamma)} & \gamma > 0,\ \kappa > 1 \ \text{ or } \ \gamma < 0,\ \kappa < 0, \\ -\langle f^{1+\gamma}\rangle^{\kappa/(1+\gamma)} & \gamma < 0,\ 0 < \kappa < 1. \end{cases}$$

For $\gamma < 0$, the reverse Minkowski inequality ensures that $-\langle f^{1+\gamma}\rangle^{1/(1+\gamma)}$ is convex in f. For $\gamma > 0$, $\kappa > 1$ or $\gamma < 0$, $\kappa < 0$, the corresponding Bregman divergence is given as

$$D(f,g) = \langle f^{1+\gamma}\rangle^{\kappa/(1+\gamma)} + \langle g^{1+\gamma}\rangle^{\kappa/(1+\gamma)} \left((\kappa - 1) - \kappa \frac{\langle f g^{\gamma}\rangle}{\langle g^{1+\gamma}\rangle} \right).$$

For the parameter $\gamma < 0$ and $0 < \kappa < 1$, the divergence is the negative of the above. The parameter $\kappa = 1 + \gamma$ yields the density-power divergence, and the parameter $\kappa = 1$ does the pseudo-spherical divergence. In this paper, this divergence is denoted as the Bregman-Hölder divergence, and the divergence with positive γ is considered in [5]. The Bregman-Hölder divergence is characterized by the intersection of Bregman divergence and Hölder divergence. This fact is proved in Theorem 3.

Example 9 (α-divergence and Hölder divergence)**.** *The α-divergence with $\alpha \neq 0, 1$ in Example 5 is represented by using the Hölder divergence, though it is not a member in the class of the composite scores. Indeed, using the density-power divergence in Example 3, we have*

$$D^{(\alpha)}_{\mathrm{alpha}}(f,g) = \frac{\alpha - 1}{\alpha} D^{(1/\alpha - 1)}_{\mathrm{power}}(f^{\alpha}, g^{\alpha})$$

for $\alpha \neq 0, 1$.

4. Theoretical Properties of Hölder Divergences

In this section, we present some theoretical properties of Hölder divergence.

4.1. Conjugate Relation

Let us consider the conjugate relation among Hölder divergences. Firstly, we point out that the KL divergence and IS distance are related to each other by the equality,

$$D_{\mathrm{IS}}(f,g) = D_{\mathrm{KL}}(1, f/g),$$

i.e., for Hölder divergence, the equality $D_{-1}(f,g) = D_0(1, f/g)$ holds. This relation is extended to Hölder divergences.

Suppose $\gamma \neq 0, -1$, and let $D^{\phi}_{\gamma}(f,g)$ be the Hölder divergence with γ and ϕ. Let $\gamma^* = -1 - \gamma$ and $\phi^*(z)$ be $z^s \phi(1/z)$, in which $s \in \{1, -1\}$ is determined from γ as shown in Definition 4. Since $\gamma^* \in (-1, 0)$ for $\gamma \in (-1, 0)$ and $\gamma^* \notin [-1, 0]$ for $\gamma \notin [-1, 0]$ hold, we define $s^* = s$. We find that $\phi(z) \geq -z^{s(1+\gamma)}$ guarantees the inequality $\phi^*(z) \geq -z^{s^*(1+\gamma^*)}$. It is straightforward to confirm that the equality

$$D^{\phi^*}_{\gamma^*}(f,g) = D^{\phi}_{\gamma}(f^{-\gamma/(1+\gamma)}, f^{1/(1+\gamma)} g^{-1}),$$

or equivalently,

$$D^{\phi^*}_{-1-\gamma}(f^{-1-1/\gamma}, f^{-1/\gamma} g^{-1}) = D^{\phi}_{\gamma}(f,g),$$

holds. Let ι be the transformation

$$\iota : (\gamma, \phi, f, g) \longrightarrow (\gamma^*, \phi^*, f^{-1-1/\gamma}, f^{-1/\gamma} g^{-1}).$$

Then, $\iota \circ \iota$ is the identity map. This implies that the Hölder divergences, D_γ^ϕ and $D_{\gamma^*}^{\phi^*}$, are connected by the conjugate relation. In the current setup, though the IS-distance D_{-1} is represented by the KL-divergence D_0, the representation of $D_0(f,g)$ by using D_{-1} is not properly defined.

4.2. Bregman Divergence and Hölder Divergence

Since the Hölder divergence $D_\gamma(f,g)$ is not necessarily convex in f, the Hölder divergence is not always represented as the form of a Bregman divergence. Let us identify the equivalence class of the intersection of Bregman divergences and Hölder divergences.

Theorem 3. *Let $D_\gamma(f,g) = S_\gamma(f,g) - S_\gamma(f,f)$ with ϕ be a Hölder divergence.*

- *If S_γ is equivalent with the score S that induces the Bregman divergence $D(f,g) = S(f,g) - S(f,f)$. Then, $D(f,g)$ is the Bregman-Hölder divergence in Example 8.*
- *If S_γ is equivalent with the score S that induces the separable Bregman divergence $D(f,g) = S(f,g) - S(f,f)$. Then, $D(f,g)$ is the density-power divergence.*

For $\gamma > 0$, the theorem was proved in [5]. We present the proof for $\gamma < 0$. The proof is found in Appendix A.

Amari [28] studied the intersection between Bregman divergence and Csiszár f-divergence under the power-representation of probability distributions. There are some attempts to define the divergence that connects the density-power divergence and the pseudo-spherical divergence [22]. The Bregman-Hölder divergence is different from the existing one.

4.3. Characterization of Hölder Scores

In Section 3, we showed that the Hölder divergence is defined from the composite score and have the scale-invariance property. Conversely, we show that these properties characterize the class of Hölder divergences. Some technical assumptions are introduced in the below.

Assumption 1. *Let $D(f,g) = S(f,g) - S(f,f)$ be the divergence for the positive-valued functions f, g on the compact support Ω.*

(a) *$D(f,g)$ satisfies the scale-invariance property Equation (2), and $S(f,g)$ is expressed as the composite score $\psi(\langle fU(g)\rangle, \langle V(g)\rangle)$.*

(b) *The functions U, V and ψ are differentiable. The two-dimensional gradient vector of ψ does not vanish on any point $x \in \mathbb{R}^2$ that is expressed as $x = (\langle fU(f)\rangle, \langle V(f)\rangle)$ for a positive-valued function f.*

Theorem 4. *Suppose that the divergence $D(f,g) = S(f,g) - S(f,f)$ satisfies Assumption 1. Then, the composite score $S(f,g)$ is equivalent with the Hölder score.*

We use the following lemmas to prove Theorem 4.

Lemma 1. *Suppose that $D(f,g)$ is the divergence defined from the composite score, $S(f,g) = \psi(\langle fU(g)\rangle, \langle V(g)\rangle)$. We assume the condition (b) in Assumption 1. Then, $V'(z) = czU'(z)$ holds with a non-zero constant $c \in \mathbb{R}$.*

Lemma 2. *Under Assumption 1, the functions $U(z)$ and $V(z)$ are given by one of followings:*

- *$U(z) = z^\gamma + c$ and $V(z) = z^{1+\gamma}$ for $\gamma \neq 0, -1$.*
- *$U(z) = -\log z + c$ and $V(z) = z$ and $c \in \mathbb{R}$.*
- *$U(z) = 1/z + c$ and $V(z) = \log z$ and $c \in \mathbb{R}$.*

The proof of Lemma 1 is shown in Lemma C.1 of [5], and hence, we omit the proof. Lemma 2 for positive γ is also proved in [5] under slightly different conditions. The proof of Lemma 2 is shown in Appendix B. Some involved argument is required to specify the expression of the function ψ of the composite score. The detailed proofs are found in Appendix C.

For the probability densities f and g defined on a non-compact support $\mathbb{R}^d$, Kanamori and Fujisawa [5] specified the expression of divergence $D(f,g)$ having the affine invariance for the coordinate x. In such case, the Hölder divergence with negative γ such as the Itakura-Saito distance is excluded, since they are not defined for functions on the non-compact support. In the present paper, we consider the divergences for the positive-valued functions on the compact support Ω.

Separable Bregman divergences are derived from composite scores. Hence, we obtain the following result.

Corollary 5. *Suppose that the separable Bregman divergence is scale-invariant. Then, the divergence should be the density-power divergence.*

Different invariance property provides different divergences. Indeed, the invariance under any invertible and differentiable data transformation leads to the Csiszár φ-divergence [7,29], and a different type of the scale-invariance leads to the pseudo-spherical divergence [24].

5. Conclusions

We proposed Hölder divergence as defined from the composite score, and showed that the Hölder divergence has the scale-invariance property. In addition, we proved that the composite score satisfying the scale-invariance property leads to the Hölder divergence. Hölder divergence is determined by a real number γ and a function ϕ. In the previous work [5], the Hölder divergence with a positive γ was proposed from the affine-invariance, and it was used to the robust parameter estimation. In this paper, we extended the previous work to Hölder divergence, having even negative parameter γ. As a result, the density-power divergence with a negative parameter and Itakura-Saito distance were unified under the Hölder divergence. The Hölder divergence with a non-negative γ can be used to measure the discrepancy between two non-negative functions on a non-compact support. On the other hand, the Hölder divergence defined from any real number γ is available to measure the degree of nearness between two non-negative functions on a compact domain. Technically, the reverse Hölder inequality and the reverse Minkowski inequality were used to prove the non-negativity of the divergence. Functions with a

compact support are also useful in statistical data analysis, though most of frequently-used densities are defined on non-compact set such as the normal distribution. Indeed, the power spectrum densities are defined on the compact set $[-\pi, \pi]$, and the IS-distance is used to measure the discrepancy between two power spectrum densities.

We presented a method of constructing the scale-invariant divergences from the (reverse) Hölder inequality. This is a new approach for introducing a class of divergences. We expect that the new class of divergences open up a new applications in the field of information sciences.

Acknowledgments

Takafumi Kanamori was partially supported by JSPS KAKENHI Grant Number 24500340.

Conflicts of Interest

The authors declare no conflict of interest.

Appendix

A. Proof of Theorem 3

Proof of case 1. Let $G(g)$ be the potential of the Bregman divergence $D(f, g)$. Suppose that there exists a strictly monotone increasing function ξ such that

$$-G(g) - \langle G_g^*(f-g)\rangle = \xi(S_\gamma(f,g)) \tag{4}$$

holds for the Hölder score S_γ. Substituting f into g, we have $G(g) = -\xi(-\langle g^{1+\gamma}\rangle)$ for $\gamma \notin [-1, 0]$ and $G(g) = -\xi(-1/\langle g^{1+\gamma}\rangle)$ for $\gamma \in (-1, 0)$. We prove the case of $\gamma \notin [-1, 0]$. The same proof works for the other case. Let $x = \langle g^{1+\gamma}\rangle$ and $z = \langle fg^\gamma\rangle/\langle g^{1+\gamma}\rangle$. Then, the Equation (4) is rewritten as

$$-\xi(-x) - (1+\gamma)\xi'(-x)(xz - x) = \xi(\phi(z)x).$$

By differentiating the both sides twice by z and setting $z = 1$, we have

$$x\xi'(-x)\phi''(1) + x^2(\phi'(1))^2\xi''(-x) = 0.$$

The solution of the differential equation is given by $\xi(x) = c_0 + c_1 x^\alpha$, where c_0, c_1, α are constants. Hence, the potential is represented as $G(f) = c\langle f^{1+\gamma}\rangle^{\kappa/(1+\gamma)}$ where c and κ are constants. Due to the convexity of $G(f)$, the parameters c and κ are determined as shown in Example 8. □

Proof of case 2. Since $S(f, g)$ and $S_\gamma(f, g)$ is equivalent, there exists a monotone function ξ such that

$$-\langle J(g) + J'(g)(f-g)\rangle = \xi\left(\phi\left(\frac{\langle fg^\gamma\rangle}{\langle g^{1+\gamma}\rangle}\right)\langle g^{1+\gamma}\rangle^s\right)$$

holds, where $s \in \{1, -1\}$. By setting $g = f$, we obtain $\xi(\langle f^{1+\gamma}\rangle^s) = -\langle J(f)\rangle$. Setting f to be a constant function $f(x) = a, x \in [0, 1]$, we obtain $J(a) = -\xi(a^{(1+\gamma)s})$. Thus, the equality

$$\xi(\langle f^{1+\gamma}\rangle^s) = \langle \xi(f^{(1+\gamma)s})\rangle$$

should hold. Let $f(x)$ on $[0,1]$ be the step function defined as $f(x) = a > 0$ for $0 \leq x \leq p$ and $f(x) = b > 0$ for $p < x \leq 1$, where $p \in (0,1)$. Then, the equality

$$\xi(\{pa^{1+\gamma} + (1-p)b^{1+\gamma}\}^s) = p\xi(a^{(1+\gamma)s}) + (1-p)\xi(b^{(1+\gamma)s})$$

holds for all p, a, b. This implies that $\xi(z^s)$ is an affine function with respect to $z > 0$. Therefore, we obtain $\langle J(f)\rangle = c_0 + c_1\langle f^{1+\gamma}\rangle$. Due to the convexity of $\langle J(f)\rangle$ in f, we find $c_1 > 0$ for $-1 > \gamma$ and $c_1 < 0$ for $0 > \gamma > -1$. As the result, only the separable Bregman score defined from $J(z) = c_1 z^{1+\gamma}$ is equivalent with the Hölder score. This is nothing but the density-power score extended to the negative parameter γ. □

B. Proof of Lemma 2

Suppose f and g be positive-valued functions defined on $\Omega = [0,1]$. Extension to the compact set in the multi-dimensional space is straightforward.

Proof. Let us consider the transformation $f(x) \mapsto f_{a,1}(x) = af(x)$ for a positive real number $a > 0$. The scale-invariance property (2) leads to

$$h(a)D(f_{a,1}, g_{a,1}) = D(f, g),$$

where $h(a) = 1/\kappa(a,1)$. Let $f(x)$ be a constant function, $f(x) = 1$ for $x \in [0,1]$, and $v(x)$ be a function such that $\sup_{x\in[0,1]} |v(x)| < 1$. For any ε such that $|\varepsilon| < 1$, the scale-invariance property leads to

$$\frac{\partial}{\partial a} h(a)D((f+\varepsilon v)_{a,1}, g_{a,1}) = 0.$$

Therefore, we obtain

$$\frac{\partial^2}{\partial\varepsilon\partial a} h(a)D((f+\varepsilon v)_{a,1}, g_{a,1})\Big|_{a=1,\varepsilon=0} = 0.$$

Some algebra yields that the above equation is expressed as

$$\int_0^1 \{c_0 + c_1 U(g(x)) + c_2 g(x)U'(g(x))\}v(x)dx = 0$$

for any function $v(x)$, where c_0, c_1 and c_2 are constants. Hence, we have

$$c_0 + c_1 U(g(x)) + c_2 g(x)U'(g(x)) = 0.$$

for any positive-valued function $g(x)$. As a result, the function $U(z)$ should satisfy the differential equation

$$c_0 + c_1 U(z) + c_2 zU'(z) = 0$$

for $z > 0$. Up to a constant factor, the solution is given as $U(z) = z^\gamma + c$ or $U(z) = \log z + c$ for $\gamma, c \in \mathbb{R}$. Due to Lemma 1, we have $V(z) = z^{1+\gamma}$ for $U(z) = z^\gamma + c$ with $\gamma \neq 0, -1$, $V(z) = \log z$ for $U(z) = 1/z + c$, and $V(z) = z$ for $U(z) = -\log z + c$ up to a constant factor. As shown in [5], the relative invariance under the transformation $f(x) \mapsto f_{1,\sigma}(x) = \sigma f(\sigma x)$ provides the same solution. □

C. Proofs of Theorem 4

Proof for $U(z) = -\log z + c$ *and* $V(z) = z$. The composite score is given as

$$S(f, g) = \psi(\langle -f \log g + cf \rangle, \langle g \rangle).$$

For any pair of positive functions f, g, the inequality $\langle -f \log g + cf \rangle + \langle g \rangle \geq \langle -f \log f + cf \rangle + \langle f \rangle$ holds, and the equality holds if and only if $f = g$. Hence, for the function ψ, the equality $\psi(x, y) = \psi(z, w)$ holds for $x+y = z+w$, and the inequality $\psi(x, y) > \psi(z, w)$ holds for $x+y > z+w$. Therefore, $\psi(x, y)$ is expressed as $\xi(x+y)$ by using a strictly increasing function ξ. As a result, the score is given as $S(f, g) = \xi(\langle -f \log g + g + cf \rangle)$ that is equivalent with the Hölder score with $\gamma = 0$ up to a monotone transformation. □

Proof for $U(z) = 1/z + c$ *and* $V(z) = \log z$. The composite score is given as

$$S(f, g) = \psi(\langle f/g + cf \rangle, \langle \log g \rangle).$$

Remember that $\langle f/g + \log g \rangle - \langle 1 + \log f \rangle$ is nothing but the Itakura-Saito distance. Hence, for any pair of positive functions f, g, the inequality $\langle f/g + cf \rangle + \langle \log g \rangle \geq \langle 1 + cf \rangle + \langle \log f \rangle$ holds, and the equality holds if and only if $f = g$. Hence, for the function ψ, the equality $\psi(x, y) = \psi(z, w)$ holds for $x + y = z + w$, and the inequality $\psi(x, y) > \psi(z, w)$ holds for $x + y > z + w$. Therefore, $\psi(x, y)$ is expressed as $\xi(x + y)$ by using a strictly increasing function ξ. The score should be represented as $S(f, g) = \xi(\langle f/g + cf + \log g \rangle)$, that is equivalent with the Hölder score with $\gamma = -1$. □

In the above proof, the identity function $\xi(z) = z$ leads to the Itakura-Saito distance, and $\xi(z) = e^z$ with $c = 0$ also leads to another scale-invariant divergence.

Proof for $U(z) = z^\gamma + c$ *and* $V(z) = z^{1+\gamma}$ *with* $\gamma \neq 0, -1$. In the paper [5], the case of $\gamma > 0$ is proved. Lemma C.2 of [5] showed that the scale-invariance of $p \mapsto p_{1,\sigma}$, leads to $\psi(x, y) = \phi((x-c)/y)y^s$, where $\phi : \mathbb{R}_+ \to \mathbb{R}$ and $s \in \mathbb{R}$. Even for $0 > \gamma \neq -1$, we find that the proof works. As a result, we see that the score with $0 > \gamma \neq -1$ is expressed as

$$S(f, g) = \phi\left(\frac{\langle f(g^\gamma + c) \rangle - c\langle f \rangle}{\langle g^{1+\gamma} \rangle}\right) \langle g^{1+\gamma} \rangle^s = \phi\left(\frac{\langle fg^\gamma \rangle}{\langle g^{1+\gamma} \rangle}\right) \langle g^{1+\gamma} \rangle^s.$$

We prove that the above $S(f, g)$ is equivalent with the Hölder score with $\gamma < 0$. Let g be $g(x) = 1$ for $x \in \Omega = [0, 1]$. Then, $S(f, g) - S(f, f) = \phi(\langle f \rangle) - \phi(1)\langle f^{1+\gamma} \rangle^s$. If $s = 0$ or $\phi(1) = 0$ holds, the equality $S(f, g) - S(f, f) = 0$ holds for any f such that $\langle f \rangle = 1$. This is the contradiction. Hence, we have $s\phi(1) \neq 0$. Hence, it is sufficient to consider the case of $s = \pm 1$ as the representative of the equivalent class.

Let us consider the sign of $s\phi(1)$. Since $S(f, g)$ defines the divergence $D(f, g)$, the inequality

$$S(f, g) - S(f, f) = \left\{\phi\left(\frac{\langle fg^\gamma \rangle}{\langle g^{1+\gamma} \rangle}\right) \frac{\langle g^{1+\gamma} \rangle^s}{\langle f^{1+\gamma} \rangle^s} - \phi(1)\right\} \langle f^{1+\gamma} \rangle^s \geq 0$$

holds. Let f and g be functions such that

$$1 = \frac{\langle fg^\gamma \rangle}{\langle g^{1+\gamma} \rangle} > \left(\frac{\langle f^{1+\gamma} \rangle}{\langle g^{1+\gamma} \rangle}\right)^{1/(1+\gamma)} \tag{5}$$

for $\gamma < 0$, *i.e.*, the reverse Hölder's inequality strictly holds for f and g such that $1 = \langle fg^\gamma\rangle/\langle g^{1+\gamma}\rangle$. Such choice is possible. For example, for linearly independent functions f and g_0 with $\langle fg_0^\gamma\rangle \neq 0$, let g be $g_0\langle fg_0^\gamma\rangle/\langle g_0^{1+\gamma}\rangle$. For $\gamma \in (-1, 0)$, we have $1 < \langle g^{1+\gamma}\rangle/\langle f^{1+\gamma}\rangle$, and for $\gamma < -1$, we have $0 < \langle g^{1+\gamma}\rangle/\langle f^{1+\gamma}\rangle < 1$. For such f and g, the inequality

$$\phi\left(\frac{\langle fg^\gamma\rangle}{\langle g^{1+\gamma}\rangle}\right)\frac{\langle g^{1+\gamma}\rangle^s}{\langle f^{1+\gamma}\rangle^s} - \phi(1) = \phi(1)\left(\frac{\langle g^{1+\gamma}\rangle^s}{\langle f^{1+\gamma}\rangle^s} - 1\right) \geq 0.$$

should hold. As a result, we have $\phi(1)s > 0$ for $\gamma \in (-1, 0)$ and $\phi(1)s < 0$ for $\gamma < -1$.

We prove that $S(f, g) = \phi(\langle fg^\gamma\rangle/\langle g^{1+\gamma}\rangle)\langle g^{1+\gamma}\rangle^s$ with $\gamma < 0$ and $s = \pm 1$ leads to the divergence $D(f, g) = S(f, g) - S(f, f)$ only when $\phi(z)$ enjoys $(1 + \gamma)s\phi(1) > 0$ and $\phi(z) \geq \phi(1)z^{(1+\gamma)s}$. The inequality $(1 + \gamma)s\phi(1) > 0$ was proved in the above. Suppose that there exists $z_0 > 0$ such that $\phi(z_0) < \phi(1)z_0^{(1+\gamma)s}$. Choose f and g such that

$$\left(\frac{\langle fg^\gamma\rangle}{\langle g^{1+\gamma}\rangle}\right)^{1+\gamma} = \frac{\langle f^{1+\gamma}\rangle}{\langle g^{1+\gamma}\rangle} = z_0^{1+\gamma}$$

holds. This is possible by choosing, say, $g = f/z_0$ for some f. For such f and g, we have

$$\begin{aligned} S(f, g) - S(f, f) &= \phi(z_0)\langle g^{1+\gamma}\rangle^s - \phi(1)\langle f^{1+\gamma}\rangle^s \\ &< \phi(1)z_0^{(1+\gamma)s}\langle g^{1+\gamma}\rangle^s - \phi(1)\langle f^{1+\gamma}\rangle^s \\ &= \phi(1)\frac{\langle f^{1+\gamma}\rangle^s}{\langle g^{1+\gamma}\rangle^s}\langle g^{1+\gamma}\rangle^s - \phi(1)\langle f^{1+\gamma}\rangle^s \\ &= 0, \end{aligned}$$

in which $\langle g^{1+\gamma}\rangle > 0$ is used. This is the contradiction. Therefore, the inequality $\phi(z) \geq \phi(1)z^{(1+\gamma)s}$ should hold for all $z > 0$.

For $\gamma < -1$ and $s = -1$, we have $\phi(1) > 0$, and we obtain $\phi(z) \geq \phi(1)z^{-(1+\gamma)} \geq 0$ for $z \geq 0$. the score $S_\gamma(f, g)$ with $s = -1$ and $\phi(z)$ is equivalent with the score $S_\gamma(f, g)$ with $s = 1$ and $-1/\phi(z) \geq -z^{1+\gamma}/\phi(1)$. As a result, for the score with $\gamma < -1$, the parameter s can be fixed to $s = 1$. For $0 > \phi(1) = -1$, the score $S(f, g)$ is equivalent with the Hölder score with the same γ.

In the same way, For $\gamma \in (-1, 0)$ and $s = 1$, we have $\phi(1) > 0$ due to $(1 + \gamma)s\phi(1) > 0$. Then, we obtain $\phi(z) \geq \phi(1)z^{1+\gamma} \geq 0$ for $z \geq 0$. Hence, for $\gamma \in (-1, 0)$, the score $S_\gamma(f, g)$ with $s = 1$ and $\phi(z)$ is equivalent with the score $S_\gamma(f, g)$ with $s = -1$ and $-1/\phi(z) \geq -z^{-(1+\gamma)}/\phi(1)$. Therefore, for the score with $\gamma \in (-1, 0)$, the parameter s can be fixed to $s = -1$. When $0 > \phi(1) = -1$ holds, the score $S(f, g)$ is equivalent with the Hölder score with $\gamma \in (-1, 0)$. □

References

1. Bremnes, B.J. Probabilistic forecasts of precipitation in terms of quantiles using NWP model output. *Mon. Weather Rev.* **2004**, *132*, 338–347.
2. Brier, G.W. Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.* **1950**, *78*, 1–3.
3. Duffie, D.; Pan, J. An overview of value at risk. *J. Deriv.* **1997**, *4*, 749.

4. Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86.
5. Kanamori, T.; Fujisawa, H. Affine invariant divergences associated with composite scores and its applications. *Bernoulli* **2014**, in press.
6. Gneiting, T.; Raftery, A.E. Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.* **2007**, *102*, 359–378.
7. Qiao, Y.; Minematsu, N. A study on invariance of f-divergence and its application to speech recognition. *IEEE Trans. Signal Process.* **2010**, *58*, 3884–3890.
8. Dawid, A.P.; Lauritzen, S.; Parry, M. Proper local scoring rules on discrete sample spaces. *Ann. Stat.* **2012**, *40*, 593–608.
9. Parry, M.; Dawid, A.P.; Lauritzen, S. Proper local scoring rules. *Ann. Stat.* **2012**, *40*, 561–592.
10. Hendrickson, A.D.; Buehler, R.J. Proper scores for probability forecasters. *Ann. Math. Stat.* **1971**, *42*, 1916–1921.
11. Bregman, L.M. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex prog ramming. *USSR Comput. Math. Math. Phys.* **1967**, *7*, 200–217.
12. Ali, S.M.; Silvey, S.D. A general class of coefficients of divergence of one distribution from another. *J. R. Stat. Soc. Ser. B* **1966**, *28*, 131–142.
13. Csiszár, I. Information-type measures of difference of probability distributions and indirect observation. *Stud. Sci. Math. Hung.* **1967**, *2*, 229–318.
14. Borwein, J.M.; Zhu, Q.Q.J. *Techniques of Variational Analysis*; CMS Books in Mathematics; Springer Science + Business Media, Incorporated: New York, NY, USA, 2005.
15. Murata, N.; Takenouchi, T.; Kanamori, T.; Eguchi, S. Information geometry of U-boost and bregman divergence. *Neural Comput.* **2004**, *16*, 1437–1481.
16. Collins, M.; Schapire, R.E.; Singer, Y. Logistic regression, adaBoost and bregman distances. In Proceedings of the Thirteenth Annual Conference on Computational Learning Theory, Palo Alto, CA, USA, 28 June–1 July 2000; pp. 158–169.
17. Banerjee, A.; Merugu, S.; Dhillon, I.S.; Ghosh, J. Clustering with bregman divergences. *J. Mach. Learn. Res.* **2005**, *6*, 1705–1749.
18. Itakura, F.; Saito, S. Analysis synthesis telephony based on the maximum likelihood method. In Proceedings of the 6th International Congress on Acoustics, Tokyo, Japan, 21–28 August 1968; pp. 17–20.
19. Févotte, C.; Bertin, N.; Durrieu, J.L. Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis. *Neural Comput.* **2009**, *21*, 793–830.
20. Fevotte, C.; Cemgil, A.T. Nonnegative matrix factorisations as probabilistic inference in composite models. In Proceedings of the 17th European Signal Processing Conference (EUSIPCO'09), Glasgow, Scotland, 24–28 August 2009.
21. Basu, A.; Harris, I.R.; Hjort, N.L.; Jones, M.C. Robust and efficient estimation by minimising a density power divergence. *Biometrika* **1998**, *85*, 549–559.
22. Jones, M.C.; Hjort, N.L.; Harris, I.R.; Basu, A. A comparison of related density-based minimum divergence estimators. *Biometrika* **2001**, *88*, 865–873.

23. Good, I.J. *Comment on "Measuring Information and Uncertainty," by R. J. Buehler*; Godambe, V.P., Sprott, D.A., Eds.; Foundations of Statistical Inference: Toronto, Canada, 1971; pp. 337–339.
24. Fujisawa, H.; Eguchi, S. Robust parameter estimation with a small bias against heavy contamination. *J. Multivar. Anal.* **2008**, *99*, 2053–2081.
25. Amari, S.; Nagaoka, H. *Methods of Information Geometry: Translations of Mathematical Monographs*; Oxford University Press: Providence, RI, USA, 2000; Volume 191.
26. Cichocki, A.; Amari, S. Families of alpha- beta- and gamma- divergences: Flexible and robust measures of similarities. *Entropy* **2010**, *12*, 1532–1568.
27. Berger, J.O. *Statistical Decision Theory and Bayesian Analysis*; Springer Series in Statistics; Springer: New York, NY, USA, 1985.
28. Amari, S. Alpha-divergence is unique, belonging to both f-divergence and Bregman divergence classes. *IEEE Trans. Inf. Theory* **2009**, *55*, 4925–4931.
29. Pardo, M.C.; Vajda, I. About distances of discrete distributions satisfying the data processing theorem of information theory. *IEEE Trans. Inf. Theory* **1997**, *43*, 1288–1293.