Affine Invariant Divergences associated with Composite Scores and its Applications

Takafumi Kanamori  
Nagoya University  
kanamori@is.nagoya-u.ac.jp

Hironori Fujisawa  
The Institute of Statistical Mathematics  
fujisawa@ism.ac.jp

May 14, 2013

Abstract

In statistical analysis, measuring a score of predictive performance is an important task. In many scientific fields, appropriate scores were tailored to tackle the problems at hand. A proper score is a popular tool to obtain statistically consistent forecasts. Furthermore, a mathematical characterization of the proper score was studied. As a result, it was revealed that the proper score corresponds to a Bregman divergence, which is an extension of the squared distance over the set of probability distributions. In the present paper, we introduce composite scores as an extension of the typical scores in order to obtain a wider class of probabilistic forecasting. Then, we propose a class of composite scores, named Hölder scores, that induce equivariant estimators. The equivariant estimators have a favorable property, implying that the estimator is transformed in a consistent way, when the data is transformed. In particular, we deal with the affine transformation of the data. By using the equivariant estimators under the affine transformation, one can obtain estimators that do not essentially depend on the choice of the system of units in the measurement. Conversely, we prove that the Hölder score is characterized by the invariance property under the affine transformations. Furthermore, we investigate statistical properties of the estimators using Hölder scores for the statistical problems including estimation of regression functions and robust parameter estimation, and illustrate the usefulness of the newly introduced scores for statistical forecasting.

keywords: composite score; divergence; Bregman score; Hölder score; affine invariance.

1 Introduction

In statistical analysis, an important task is to measure a score or a loss of the prediction performance. In many fields in which probabilistic forecasting is required, appropriate scoring rules or loss functions are tailored to tackle the scientific problems at hand, e.g., weather and climate prediction \cite{8,9}, computational finance \cite{15}, and so forth.

Under an uncertain situation, the prediction is described by using the probability distribution. The probability distribution for the prediction is expected to put much weight to outcomes that are likely to materialize in the future. Hence, the score is formalized as a function taking two inputs, i.e., a probability distribution for the prediction and an outcome. In order to achieve high prediction performance on average, ideally, optimization of the expected score is
conducted. When the identically and independently distributed (i.i.d.) samples are available, the expected score is approximated by the empirical mean over the samples. By optimizing the empirical mean of the score over a statistical model for the prediction, one will obtain a probability distribution attaining high prediction performance.

The above statistical procedure is formalized as the statistical inference using the scores or scoring rules [9, 13, 20, 24]. We regard the score as a loss to be minimized. The estimator obtained from the score is called the optimum score estimator. To obtain a good estimator, scores need to satisfy some assumptions. A typical assumption is that the score is proper. Given a probability distribution of outcomes, the optimal value of the expected proper score is attained by setting the prediction probability to be the true probability distribution. Under mild assumptions, optimization of the proper score averaged over the observed samples produces a statistically consistent estimator. The proper score is a special case of M-estimation [25], and the statistical property of the proper score have been studied in the framework of M-estimators [32, Chap. 5].

The proper score is a basic element that yields important concepts in statistical inference. According to [13], the proper score introduces a divergence, which is a discrepancy measure between two probability distributions. The divergence is regarded as a generalization of the (squared) distance, and induces a sort of topological structure over the statistical model. As a result, geometrical structures such as the Riemannian metric and affine connection are defined over the geometrical space consisting of probability distributions. Such a geometrical structure is closely related to the statistical properties of the estimator using the corresponding proper score. Bregman divergence [7] is an important class of divergences, since it is closely related to the proper score. A major milestone in the theoretical approach is the characterization of the proper score by using the Bregman divergence [1, 20, 24]. More precisely, any proper score produces a Bregman divergence, and a given Bregman divergence yields a proper score. The correspondence established a way to investigate the proper score by using the Bregman divergence on statistical models.

In the present paper, we introduce composite scores as an extension of the proper scores in order to obtain a wider class of probabilistic forecasting. Then, we propose a class of composite scores, named Hölder scores, that induce equivariant estimators [5]. The equivariant estimator is a class of estimators having a favorable property, implying that the estimator is transformed in a consistent way, when the data is transformed. In particular, we deal with the affine transformation of the data, i.e., \( \omega \mapsto \sigma^{-1}(\omega - \mu) \) for the data \( \omega \in \mathbb{R}^d \), where \( \sigma \) is a \( d \times d \) invertible matrix and \( \mu \) is a \( d \)-dimensional vector. The normalization of data is a typical example of affine transformations. Each element of the normalized data has zero sample-mean and unit sample-variance. Thus, for the normalized data, the statistical comparison of each component is reasonable. As an example of the equivariant estimators under the affine transformation, let us consider the estimation of the mean value \( \theta \) of a one-dimensional probability distribution. When all samples are transformed from \( \omega \in \mathbb{R} \) into \( \sigma^{-1}(\omega - \mu) \) with the constants \( \mu \in \mathbb{R} \) and \( \sigma \neq 0 \), also the estimator \( \hat{\theta} \) of the mean value \( \theta \) should be transformed into \( \sigma^{-1}(\hat{\theta} - \mu) \). By using the equivariant estimators under the affine transformation, the estimate does not essentially depend on the choice of the system of units in the measurement. In addition, we show a characterization of the Hölder scores. Similarly to the correspondence between the proper scores and the Bregman divergences, the composite scores correspond to a class of divergences. When the divergence is invariant under the data transformation, the corresponding composite score provides an equivariant estimator. We prove that the Hölder score is characterized by the affine invariance of the associated divergence, i.e., among a class of composite scores, only
Hölder score provides the equivariant estimator under affine transformations. Furthermore, we investigate statistical properties of the estimators derived from Hölder scores for the statistical problems including estimation of regression functions and robust parameter estimation.

As pointed out in [8], scores of continuous variables have so far received little attention. In this paper, our main concern is the scores of continuous variables. The invariance under affine transformations is a specific property for continuous variables.

The remainder of the article is organized as follows. In Section 2, we define composite scores and associated divergences. Bregman scores and its separable variant are also introduced as an important class of composite scores. Then, we show a way to use composite scores to probabilistic forecasting. In Section 3, we define Hölder scores, and demonstrate the relation between Hölder scores and Bregman scores. In Section 4, we define the affine invariance of divergences, and show that the Hölder score induces the affine invariant divergences and equivariant estimators. Conversely, we prove that Hölder score is characterized by the affine invariance of the associated divergence. In Section 5, the Hölder score is used to statistical problems including regression problems and robust estimation. In particular, the robustness property of the Hölder score is presented. In Section 6, we close this article with a discussion of the possibility of the newly introduced class of scores.

2 Composite Scores and Associated Divergences

In this section, we define composite scores and associated divergences. Then, we introduce estimators using the composite scores.

Let us summarize the notations to be used throughout the paper. Let \( \mathbb{R} \) be the set of all real numbers. The non-negative numbers are denoted as \( \mathbb{R}_+ = \{ x \in \mathbb{R} \mid x \geq 0 \} \). The interior set of a set \( A \) is denoted as \( A^\circ \). Thus, \( \mathbb{R}_+^\circ \) implies the set of all positive real numbers, i.e., \( \mathbb{R}_+^\circ = \{ x \in \mathbb{R} \mid x > 0 \} \). For a sample space \( \Omega \), let \( \mathcal{B} \) be a \( \sigma \)-algebra of subsets of \( \Omega \) and \( m : \mathcal{B} \to \mathbb{R}_+ \) be a \( \sigma \)-finite measure on \( (\Omega, \mathcal{B}) \). The set of all measurable functions on \( \Omega \) is denoted as \( L_0, \) i.e., \( L_0 = \{ f : \Omega \to \mathbb{R} \mid f \text{ is measurable on } (\Omega, \mathcal{B}, m) \} \). For \( f \in L_0 \), the integral \( \int_{\Omega} f(\omega)dm(\omega) \) is denoted as \( \langle f \rangle \). Let \( \| \cdot \|_\alpha \) for \( 1 \leq \alpha < \infty \) be the \( L_\alpha \)-norm, i.e., \( \| f \|_\alpha = (\langle |f|^\alpha \rangle)^{1/\alpha} \), and \( \| \cdot \|_\infty \) be the essential sup-norm. For \( \alpha \geq 1 \), let \( L_\alpha \) be \( L_\alpha = \{ f \in L_0 \mid \| f \|_\alpha < \infty \} \). For \( \alpha = 0 \) or \( \alpha \geq 1 \), \( L_\alpha^+ \) denotes the set of all non-negative and non-zero functions in \( L_\alpha \), i.e., \( L_\alpha^+ = \{ f \in L_\alpha \mid f \geq 0, f \neq 0 \} \). Provided a set of measurable and non-negative functions \( \mathcal{F} \subset L_0^+ \), \( \mathcal{P} \) denotes the set of probability densities in \( \mathcal{F} \), i.e., \( \mathcal{P} = \{ p \in \mathcal{F} \mid \langle p \rangle = 1 \} \). For a differentiable function \( \psi \), \( \psi_i \) with the integer \( i \) denotes the partial derivative of \( \psi \) with respect to the \( i \)-th argument, e.g., for \( \psi(x, y) \), \( \psi_1 \) and \( \psi_2 \) denote \( \partial \psi / \partial x \) and \( \partial \psi / \partial y \), respectively.

2.1 Definitions

Let us consider the probabilistic forecasts on a measurable space \( (\Omega, \mathcal{B}, m) \). Suppose that the probabilistic forecast is given by a probability density \( q \in L_1^+ \) satisfying \( \langle q \rangle = 1 \). For an outcome \( \omega \in \Omega \), let \( S_0(\omega, q) \) be a score of the forecast using \( q \). When the probability density of the outcome is \( p \), the expected score is given as

\[
S_0(p, q) := \int_{\Omega} S_0(\omega, q)p(\omega)dm(\omega).
\]
Suppose that the expected score satisfies the inequality $S_0(p, q) \geq S_0(p, p)$. Then, the minimization of the empirical mean of $S_0(\omega, q)$ over the statistical model $q$ is expected to provide a good estimate of the probability density $p$. This approach is widely used in statistical inference. In this paper, the term score denotes the expected score $S_0(p, q)$, though typically the score denotes the function $S(\omega, g)$.

Let us define a general form of scores. It is defined not only for probability densities but also non-negative functions.

**Definition 1** (composite score). Let $F$ be a convex subset in $L_0^+$, and the set of probability densities in $F$ is denoted as $P$, i.e., $P = \{ p \in F \mid \langle p \rangle = 1 \}$. The function $S(f, g) : F \times F \to \mathbb{R}$ is called the composite score on $F$ if the following three conditions are satisfied:

1. $S(f, g)$ is of the form
   \[ S(f, g) = T \left( \int_{\Omega} S_0(\omega, g)f(\omega)dm(\omega), g \right), \]
   where $S_0 : \Omega \times F \to \mathbb{R}$ and $T : \mathbb{R} \times F \to \mathbb{R}$. The function $S_0(\cdot, g)f(\cdot)$ is assumed to be integrable for all $f, g \in F$.
2. $S(f, g) \geq S(f, f)$ for all $f, g \in F$.
3. For $p, q \in P$, $S(p, q) = S(p, p)$ implies $p = q$ (almost surely).

When the composite score $S(f, g)$ is defined only on the set of probability densities and the function $T$ is given as $T(c, g) = c$, the composite score is reduced to the expectation of a strictly proper score [13, 20, 22, 24]. Hence, the above definition is an extension of the strictly proper score. In Section 3, we propose a class of composite scores with a non-trivial $T$.

**Remark 1.** In our definition, the domain of the composite score is not necessarily a set of probability densities, but it can be a set of non-negative functions. Likewise, in [24], the strictly proper scores are characterized on the set of non-negative functions. The definition in the present paper simplifies mathematical analysis on composite scores.

**Definition 2** (divergence). Let $S$ be a composite score on $F$. Then, we call

\[ D(f, g) = S(f, g) - S(f, f), \quad f, g \in F, \]

the divergence associated with $S$.

By the definition of the composite score, the divergence $D(f, g)$ is nonnegative for all $f, g \in F$, and the equality $D(p, q) = 0$ for $p, q \in P$ implies $p = q$.

### 2.2 Bregman scores

As an important class of composite scores, we introduce a Bregman score and its separable variant. Under a mild assumption, any strictly proper score on $P$ is expressed as a Bregman score on $P$ [1 20 24].
Definition 3 (Bregman score). For a convex set $\mathcal{F} \subset L_0^+$, let us define $G : \mathcal{F} \to \mathbb{R}$ as a convex function such that $G$ is strictly convex on $\mathcal{P} = \{ p \in \mathcal{F} \mid \langle p \rangle = 1 \}$. Suppose that there exists a function $G_g^* : \Omega \to \mathbb{R}$ depending on $g \in \mathcal{F}$ such that

$$G(f) \geq G(g) + \int_{\Omega} G_g^*(\omega) f(\omega) dm(\omega) - \int_{\Omega} G_g^*(\omega) g(\omega) dm(\omega), \quad \text{for } f, g \in \mathcal{F}$$

holds, where the integrals are assumed to be finite. Then, the Bregman score $S(f, g)$ on $\mathcal{F}$ is defined as

$$S(f, g) = -G(g) - \int_{\Omega} G_g^*(\omega) f(\omega) dm(\omega) + \int_{\Omega} G_g^*(\omega) g(\omega) dm(\omega), \quad \text{for } f, g \in \mathcal{F}.$$  

The function $G$ is referred to as the potential function of the Bregman score, and it satisfies $G(f) = -S(f, f)$. The Bregman divergence is the divergence associated with the Bregman score.

From the definition, the Bregman score satisfies $S(f, g) \geq S(f, f)$ for all $f, g \in \mathcal{F}$. The strict convexity of $G$ on $\mathcal{P}$ ensures that the Bregman score satisfies the third condition of Definition 1, see Theorem 1 of [20]. The function $G_g^*$ corresponds to the subgradient of $G$ at $g \in \mathcal{F}$. The rigorous definition of $G_g^*$ requires the dual space of a Banach space in $L_0^+$. See [6, Chap. 4] for sufficient conditions of the existence of $G_g^*$. To avoid technical difficulties, we assume the existence of $G_g^*$ in the above definition. The Bregman score is represented as the composite score $\mathcal{S}$ with

$$S_0(\omega, g) = -G_g^*(\omega) \quad \text{and} \quad T(c, g) = c - G(g) + \langle G_g^* g \rangle.$$  

When the Bregman score is defined on the set of probability densities, setting $S_0(\omega, g) = -G_g^*(\omega) - G(g) + \langle G_g^* g \rangle$ and $T(c, g) = c$ is also a valid choice. This implies that the Bregman score on $\mathcal{P}$ is represented as a strictly proper score.

The separable variant of the Bregman score is defined as follows.

Definition 4 (separable Bregman score). Let $J : \mathbb{R}_+ \to \mathbb{R}$ be a strictly convex function. The Bregman score with the potential function $G(f) = \langle J(f) \rangle$ is called the separable Bregman score. The separable Bregman divergence is the divergence associated with the separable Bregman score.

The separable Bregman score is of the form

$$S(f, g) = -\langle J(g) \rangle - \langle J'(g)f \rangle + \langle J'(g)g \rangle \quad \text{for } f, g \in \mathcal{F},$$

where $J'(z)$ is the subgradient of $J$ at $z \in \mathbb{R}_+$.

We show some examples of Bregman scores and associated divergences.

Example 1 (Kullback-Leibler (KL) score). Let $\mathcal{F}$ be a subset of $L_1^+$, and suppose that $f \log g$ is integrable for all $f, g \in \mathcal{F}$. The Kullback-Leibler (KL) score is defined as

$$S(f, g) = \langle -f \log g + g \rangle, \quad f, g \in \mathcal{F},$$

which is the separable Bregman score using the function $J(z) = z \log z - z$ and the potential function $G(f) = \langle f \log f - f \rangle$. The associated divergence is called the KL divergence.
Example 2 (Density power score). Let $\mathcal{F} = L_{1+\gamma}^+$ for a given $\gamma > 0$. The density power score on $\mathcal{F}$ is defined as

$$S(f, g) = \langle g^{1+\gamma} \rangle - \frac{1+\gamma}{\gamma} \langle fg^\gamma \rangle, \quad f, g \in \mathcal{F},$$

which is the separable Bregman score with $J(z) = z^{1+\gamma}/\gamma$ and the potential function $G(f) = \langle f^{1+\gamma} \rangle/\gamma$. The integrability of $fg^\gamma$ is confirmed by Hölder’s inequality. The associated divergence is called the density power divergence $[3, 4, 26]$. When the parameter $\gamma$ in the density power divergence tends to zero, the KL-divergence is recovered.

Example 3 ($\gamma$-score; pseudospherical score). Let $\mathcal{F} = L_{1+\gamma}^+$ for a given $\gamma > 0$. The pseudospherical score $[21]$ is defined as

$$S(f, g) = -\frac{\langle fg^\gamma \rangle}{\langle g^{1+\gamma} \rangle^{1/(1+\gamma)}}, \quad f, g \in \mathcal{F},$$

which is the non-separable Bregman score with the potential function $G(f) = \langle f^{1+\gamma} \rangle^{1/(1+\gamma)} = \|f\|_{1+\gamma}$. For the pseudospherical score $S(f, g)$, the composite score $-\frac{1}{\gamma} \log(-S(f, g))$ is called the $\gamma$-score in this paper. The $\gamma$-score is proposed in $[16, 17]$, and it is used for robust parameter estimation. As the limiting case of $\gamma \to 0$, the divergence associated with the $\gamma$-score recovers KL-divergence.

2.3 Optimum score estimator

Statistical inference using the composite score $[11]$ is conducted by substituting the empirical probability and the model probability into the composite score. Provided the i.i.d. samples $\omega_1, \ldots, \omega_n$ from the probability density $p$, an empirical approximation of $S(p, q)$ for a given probability density $q$ is given as

$$S(\tilde{p}, q) = T \left( \frac{1}{n} \sum_{i=1}^{n} S_0(\omega_i, q), q \right),$$

where $\tilde{p}$ denotes the empirical probability. For a sufficiently large number of samples, $S(\tilde{p}, q)$ converges to $S(p, q)$ due to the law of large numbers. Since $S(p, q) \geq S(p, p)$ is assumed, the estimator of $p$ is obtained as the minimum solution of $S(\tilde{p}, q)$ with respect to $q$ over a statistical model. The estimator $\hat{q}$ is called the optimum score estimator $[20]$. The estimator using the strictly proper score is a special case of M-estimation $[25]$, and its statistical properties have been deeply investigated $[32]$.

Different composite scores may produce the same estimator. Let us define the equivalence class on the set of composite scores such that the composite scores in the same class provide the same estimator.

Definition 5 (equivalence of composite scores). The composite scores $S(f, g)$ and $\tilde{S}(f, g)$ on $\mathcal{F}$ are equivalent if there exists a strictly increasing function $\xi : \mathbb{R} \to \mathbb{R}$ such that $\tilde{S}(f, g) = \xi(S(f, g))$ holds for all $f, g \in \mathcal{F}$. The composite scores $S(f, g)$ and $\bar{S}(f, g)$ on $\mathcal{F}$ are equivalent in probability if there exists a strictly increasing function $\xi : \mathbb{R} \to \mathbb{R}$ such that $\tilde{S}(p, q) = \xi(S(p, q))$ holds for all probability densities $p, q \in \mathcal{P} \subset \mathcal{F}$. 
For any strictly increasing function $\xi$, the minimum solutions of $S(p,q)$ and $\xi(S(p,q))$ with respect to $q$ are the same. Hence, the composite scores that are equivalent in probability provide the same estimator. A different definition of the equivalence class was also proposed by [12, 13], in which the Bregman scores $S(p,q)$ and $\tilde{S}(p,q)$ on $\mathcal{P}$ are equivalent if there exist a positive constant $c > 0$ and a function $k : \mathcal{P} \to \mathbb{R}$ such that $\tilde{S}(p,q) = cS(p,q) + k(p)$ holds. The equivalence class in Definition 5 is more suitable for our analysis.

3 Hölder scores

In this section, we propose a class of composite scores, named Hölder scores, a part of which is not represented as the Bregman score. We investigate the relation between the Hölder scores and Bregman scores.

3.1 Definition of Hölder score

Bregman scores are widely used for statistical inference, [2] [11] [28] [31], since one can substitute the empirical probability distribution into the Bregman score. Under a regularity condition, Bregman scores produce statistically consistent estimators based on the outcomes. Especially, the density power score and $\gamma$-score are used for robust estimation [3] [19]. In this section, we propose a class of composite scores called Hölder scores that include both the density power score and $\gamma$-score. One can also substitute the empirical probability distribution into the Hölder score. As shown later, the Hölder score is not included in the class of Bregman scores, and has a relation to affine invariant estimators.

Definition 6 (Hölder score). The Hölder score with a nonnegative parameter $\gamma$ is defined as follows:

1. For a given $\gamma > 0$, let $\phi : \mathbb{R}_+ \to \mathbb{R}$ be a function such that $\phi(z) \geq -z^{1+\gamma}$ for all $z \geq 0$ and $\phi(1) = -1$ hold. Then, for $\mathcal{F} = L_{1+\gamma}^+$, the Hölder score is defined as

$$S(f,g) = \phi\left(\frac{\langle fg^\gamma \rangle}{\langle g^{1+\gamma} \rangle}\right) \langle g^{1+\gamma} \rangle, \quad f,g \in \mathcal{F}.$$ 

2. For $\gamma = 0$, the Hölder score is defined as

$$S(f,g) = \langle -f \log g + g \rangle, \quad f,g \in \mathcal{F},$$ 

where $\mathcal{F}$ is a subset of $L_1^+$ such that $f \log g$ is integrable for all $f,g \in \mathcal{F}$.

The associated divergence $D(f,g) = S(f,g) - S(f,f)$ is called the Hölder divergence.

The Hölder score with $\gamma = 0$ is nothing but the KL score. An appropriate choice of the function $\phi$ produces the composite score equivalent with the density power score or $\gamma$-score. Indeed, the Hölder score with the lower bound $\phi(z) = -z^{1+\gamma}$ is $S(f,g) = -\langle fg^\gamma \rangle^{1+\gamma}/\langle g^{1+\gamma} \rangle^\gamma$ which is equivalent with $\gamma$-score. The density power score is equivalent with the Hölder score with $\phi(z) = \gamma - (1 + \gamma)z$.

We prove the basic property that the Hölder score satisfies the condition of the composite score in Definition 4.
Theorem 1. The Hölder score is a composite score.

The proof of Theorem 1 is found in Appendix A. The Hölder score with $\gamma > 0$ is represented as the composite score (1) with $S_0(\omega, g) = g(\omega)\gamma$ and $T(c, g) = \phi(c/\langle g^{1+\gamma} \rangle)/\langle g^{1+\gamma} \rangle$. The name of Hölder score comes from the fact that Hölder’s inequality is used to prove the non-negativity of Hölder divergence. The function $S(f, f)$ is referred to as entropy. The entropy of the Hölder score is $S(f, f) = -\langle f^{1+\gamma} \rangle$, which is in agreement with the Tsallis entropy [30] up to an affine transformation.

3.2 Bregman scores and Hölder scores

Let us consider the relation between the Bregman scores and Hölder scores. We assume the differentiability for Bregman scores. The definition of the differentiability is shown below.

Definition 7 (differentiability of potential function). Let $G$ be the potential function of the Bregman score on the convex set $\mathcal{F}$. If the limit
\[
\lim_{\varepsilon \to 0} \frac{G((1-\varepsilon)f + \varepsilon g) - G(f)}{\varepsilon}
\]
exists for any $f, g \in \mathcal{F}$ such that there exists $\delta > 0$ satisfying $(1-\varepsilon)f + \varepsilon g \in \mathcal{F}$ for all $\varepsilon \in (-\delta, \delta)$, the potential function $G$ is differentiable. The corresponding Bregman score (resp. divergence) is called the differentiable Bregman score (resp. divergence).

The differentiability above makes our analysis rather simple. For non-differentiable Bregman scores, we will need more involved argument such as the convex analysis in Banach spaces. From the practical viewpoint, differentiable Bregman scores will be preferable, since the standard non-linear optimization techniques are directly applicable to obtain the optimum score estimator.

Theorem 2. Suppose that the function $\phi$ in the Hölder score is continuous on $\mathbb{R}_+^+$.

1. Suppose that the differentiable Bregman score with the potential function $G(f)$ is equivalent with the Hölder score with $\gamma > 0$. Then, $G(f)$ is given as $G(f) = \langle f^{1+\gamma} \rangle^{\kappa/(1+\gamma)}$ up to a positive constant factor, where $\kappa \geq 1$.

2. Suppose that the differentiable and separable Bregman score with the potential function $G(f)$ is equivalent with the Hölder score with $\gamma > 0$. Then, $G(f)$ is given as $G(f) = \langle f^{1+\gamma} \rangle$ up to a positive constant factor.

The proof is shown in Appendix B.

The KL score is a differentiable and separable Bregman. Hence, the intersection of (separable) Bregman score and Hölder score is the KL score or the (separable) Bregman score associated with the potential function presented in the above theorem.

For the potential function $G(f) = \langle f^{1+\gamma} \rangle^{\kappa/(1+\gamma)}$ with $\gamma > 0$ and $\kappa \geq 1$, the corresponding Bregman score is given as
\[
S(f, g) = \langle g^{1+\gamma} \rangle^{\kappa/(1+\gamma)} \left( 1 - \frac{1}{\kappa} - \frac{(fg^\gamma)}{(g^{1+\gamma})} \right).
\]
The above Bregman scores include the density power score \((\kappa = 1 + \gamma)\) and \(\gamma\)-score \((\kappa = 1)\) in each equivalent class. The Hölder score corresponding to the Bregman score \((2)\) is given by the function \(\phi(z)\) defined as

\[
\phi(z) = -\kappa^{(1+\gamma)/\kappa}|z - 1 + 1/\kappa|^{(1+\gamma)/\kappa}\text{sign}(z - 1 + 1/\kappa),
\]

where \(\text{sign}(z)\) is the sign function taking \(z/|z|\) for \(z \neq 0\) and 0 for \(z = 0\). In Section 5.1, we show a statistical interpretation of the composite scores included in the intersection of Bregman scores and Hölder scores.

4 Affine invariance of Hölder divergence

Affine transformation of the observed data is often used in statistical analysis. Let \(\Omega = \mathbb{R}^d\), \(B\) be the Borel set of \(\Omega\), and \(m\) be the Lebesgue measure on \((\Omega, B)\). The affine transformation is defined as the map \(\omega \mapsto \sigma^{-1}(\omega - \mu)\) of \(\omega \in \Omega\) with an invertible matrix \(\sigma \in \mathbb{R}^{d \times d}\) and a vector \(\mu \in \mathbb{R}^d\). The normalization is a typical example of the affine transformation. For the observed data \(\omega_1, \ldots, \omega_n \in \mathbb{R}^d\), let the vector \(\mu\) be the sample mean of the observations, and the matrix \(\sigma\) be the diagonal matrix such that the \(k\)-th diagonal element is equal to the sample-based standard deviation of the \(k\)-th component of the observed data. Then, each element of the transformed data, \(\sigma^{-1}(\omega_1 - \mu), \ldots, \sigma^{-1}(\omega_n - \mu)\), has zero sample-mean and unit sample-variance. This transformation enables the fair comparison of the intensity of each component in statistical sense. As another benefit, the normalization often makes the numerical computation stable.

The affine transformation of data, \(\omega \mapsto \sigma^{-1}(\omega - \mu)\), induces the transformation of the probability density,

\[
p(\omega) \mapsto p_{\sigma,\mu}(\omega) = |\det \sigma|p(\sigma\omega + \mu).
\]

Let \(q\) be a statistical model to estimate the probability density \(p\). Then, the statistical model for the affine transformed data is given as \(q_{\sigma,\mu}\). Let \(\hat{q}\) be the estimator of \(p\) based on the original data \(\{\omega_1, \ldots, \omega_n\}\), and \(\hat{q}_{\sigma,\mu}\) be the estimator based on the transformed data, \(\{\sigma^{-1}(\omega_1 - \mu), \ldots, \sigma^{-1}(\omega_n - \mu)\}\). It will be natural to require that the estimator is transformed in a consistent way, when the data is transformed, i.e., the equality

\[
(\hat{q})_{\sigma,\mu} = \hat{q}_{\sigma,\mu}
\]

should hold. The estimators enjoying \((4)\) do not essentially depend on the choice of the units in the measurement. In the present paper, the estimator satisfying \((4)\) is called the affine invariant estimator. In a formal mathematical description, the term equivariant estimator is used to denote the estimator that changes in a consistent way under data transformations \([5]\).

A simple way of obtaining the affine invariant estimator is to use the composite scores satisfying the equality \(S(p, q) = S(p_{\sigma,\mu}, q_{\sigma,\mu})\). However, the equality is not necessity. In the below, we introduce composite scores and associated divergences that provide the affine invariant estimator.

**Definition 8** (affine invariant divergence; affine invariant composite score). Let \(S\) be a composite score on \(\mathcal{F}\), and \(D\) be the associated divergence. The divergence \(D(f, g)\) is affine invariant if
there exists an $\mathbb{R}_+^d$-valued function $h(\sigma, \mu)$ of the invertible matrix $\sigma \in \mathbb{R}^{d \times d}$ and the vector $\mu \in \mathbb{R}^d$ such that the equality

$$h(\sigma, \mu) D(p_{\sigma, \mu}, q_{\sigma, \mu}) = D(p, q)$$

holds for any pair of probability densities $p, q \in \mathcal{P}$ and arbitrary affine transformation with $(\sigma, \mu)$. The function $h$ is called the scale function. The composite score $S$ inducing the affine invariant divergence is called the affine invariant composite score.

We briefly prove that the affine invariant composite score provides the affine invariant estimator. Let $S$ be an affine invariant composite score, and $\hat{q}$ be the optimum score estimator obtained by solving the minimization problem $\min_{q \in \mathcal{M}} S(p, q)$ on a statistical model $\mathcal{M}$. Then, the inequalities,

$$D(p, \hat{q}) \leq D(p, q) \quad \text{and} \quad D(p_{\sigma, \mu}, \langle \hat{q} \rangle_{\sigma, \mu}) \leq D(p_{\sigma, \mu}, q_{\sigma, \mu})$$

hold for all $q \in \mathcal{M}$. On the other hand, $\hat{q}_{\sigma, \mu}$ is the minimum solution of $\min_{q_{\sigma, \mu}} \{ q_{\sigma, \mu} \, | \, q \in \mathcal{M} \}$, when the model $\{ q_{\sigma, \mu} \, | \, q \in \mathcal{M} \}$ is used. Therefore, the equivariant property ([1]) holds, if the optimal solution is unique.

It is straightforward to verify that the Hölder score is affine invariant. Indeed, for the Hölder divergence $D(p, q)$ with $\gamma > 0$, we have

$$D(p_{\sigma, \mu}, q_{\sigma, \mu}) = \phi \left( \frac{\langle p_{\sigma, \mu} q_{\sigma, \mu} \rangle}{\langle q_{\sigma, \mu}^{1+\gamma} \rangle} \right) \langle q_{\sigma, \mu}^{1+\gamma} \rangle + \langle p_{\sigma, \mu}^{1+\gamma} \rangle$$

$$= \phi \left( \frac{|\det \sigma|^\gamma \langle pq \rangle}{\det \sigma |q^{1+\gamma}|} \right) \langle q^{1+\gamma} \rangle |\det \sigma|^\gamma + \langle p^{1+\gamma} \rangle |\det \sigma|^\gamma$$

Therefore, the scale function is given as $h(\sigma, \mu) = |\det \sigma|^{-\gamma}$. In the same way, we can confirm that the KL divergence is also affine invariant with the scale function $h(\sigma, \mu) = 1$. This result indicates that the optimum score estimator using Hölder score provides the affine invariant estimator.

Conversely, we prove that the Hölder score is characterized by the affine invariance. In the beginning, let us introduce some assumptions.

**Assumption 1** (basic assumption on $\Omega$ and $\mathcal{F}$). Let $\Omega = \mathbb{R}^d$, $\mathcal{B}$ be the Borel set of $\Omega$, and $m : \mathcal{B} \to \mathbb{R}_+$ be the Lebesgue measure on $(\Omega, \mathcal{B})$. The set $\mathcal{F}$ includes the following function set,

$$\mathcal{F}_0 := \left\{ f \in L_0^+ \left| \{ \omega \in \Omega \, | \, f(\omega) > 0 \} = (0, 1)^d, \text{ and there exist } a, b \in \mathbb{R} \text{ such that } 0 < a < f(\omega) < b \text{ for all } \omega \in (0, 1)^d \right. \right\},$$

i.e., $\mathcal{F}_0 \subset \mathcal{F} \subset L_0^+$ holds.

The subset $(0, 1)^d$ in the above assumption can be replaced with any subset with a finite measure. We assume the following conditions on the composite score.

**Assumption 2** (assumption on the composite score). For the composite score, we assume three conditions:

(a) The composite score \( S(f,g) \) has the form of
\[
S(f,g) = \psi((fU(g)),(V(g))) \quad \text{for all } f,g \in \mathcal{F},
\]
where \( U \) and \( V \) are real-valued functions on \( \mathbb{R}_+ \) and \( \psi \) is a function on a subset of \( \mathbb{R}^2 \), i.e., \( S \) is the composite score \( \mathbb{J} \) with \( S_0(\omega,g) = U(g(\omega)) \) and \( T(c,g) = \psi(c,(V(g))) \). For all \( f,g \in \mathcal{F} \), the functions \( fU(g) \) and \( V(g) \) are integrable.

(b) The functions \( U, V : \mathbb{R}_+ \rightarrow \mathbb{R} \) are second order continuously differentiable on \( \mathbb{R}^2_+ \), and they are not constant function on \( \mathbb{R}^2_+ \). For the function \( V \), the equality \( \lim_{z \rightarrow 0} V(z) = 0 = V(0) \) holds, and the limit \( \lim_{z \rightarrow 0} V'(z) \) exists.

(c) Let \( D_{U,V} \) and \( E_{U,V} \) be subsets of \( \mathbb{R}^2 \) defined as
\[
D_{U,V} = \{(fU(g)),(V(g)) \in \mathbb{R}^2 \mid f,g \in \mathcal{F}\},
\]
\[
E_{U,V} = \{(fU(f)),(V(f)) \in \mathbb{R}^2 \mid f \in \mathcal{F}\},
\]
respectively. For arbitrary point \( x \in D_{U,V} \), there exists an open neighbourhood of \( x \) on which \( \psi \) is second order continuously differentiable. For arbitrary point \( x \in E_{U,V} \), there exists an open neighbourhood of \( x \) on which the gradient vector \( (\psi_1, \psi_2) \) does not vanish.

All separable Bregman scores are expressed as the form of \( \mathbb{K} \). There exist Bregman scores that are not described by \( \mathbb{K} \), while Bregman scores do not cover all the composite scores \( \mathbb{J} \). The composite score of the form \( \mathbb{J} \) is useful in practice, since it can be calculated via integrals. In Assumption \( \mathbb{I} \), we assumed \( V(0) = 0 \) in order to guarantee the integrability of the function whose support is not equal to \( \Omega \). More precisely, let \( Z = \{\omega \in \mathbb{R}^d \mid f(\omega) = 0\} \) with \( m(Z) = \infty \), then \( \langle V(g) \rangle = \int_Z V(0)dm + \int_{\Omega \setminus Z} V(g)dm \) will not be finite unless \( V(0) = 0 \). In Assumption \( \mathbb{I} \), we assumed that the gradient vector \( (\psi_1, \psi_2) \) does not become the zero vector at \( \langle (fU(f)),(V(f)) \rangle \). If this assumption does not hold, we need a more involved argument to derive analytic properties of the functions \( U \) and \( V \). For the sake of simplicity, we introduce Assumption \( \mathbb{I} \).

The functions \( U \) and \( V \) of the affine invariant composite scores are determined by Theorem \( \mathbb{K} \).

**Theorem 3.** Let \( S \) be a composite score that produces an affine invariant divergence. Suppose that Assumption \( \mathbb{I} \) and Assumption \( \mathbb{I} \) hold. Then, the functions \( U \) and \( V \) in \( \mathbb{I} \) are given as \( U(z) = z^{\gamma} + c \) and \( V(z) = z^{1+\gamma} \) with \( \gamma > 0 \), or \( U(z) = -\log z + c \) and \( V(z) = z \) up to a constant factor, where \( c \in \mathbb{R} \) is a constant.

The proof is found in Appendix \( \mathbb{C.1} \). For each possibility of \( U \) and \( V \), the composite score is identified in the following theorem.

**Theorem 4.** Let \( S \) be a composite score that produces an affine invariant divergence. Suppose that Assumption \( \mathbb{I} \) and Assumption \( \mathbb{I} \) hold.

1. Let us define \( U(z) = -\log z + c \) and \( V(z) = z \) in \( \mathbb{I} \). Then, the composite score \( S(f,g) \) is equivalent in probability with the KL score.

2. For \( \gamma > 0 \), let us define \( U(z) = z^{\gamma} + c \) and \( V(z) = z^{1+\gamma} \) in \( \mathbb{I} \), and let \( \mathcal{F} \) be \( \mathcal{F} = L_{1+\gamma}^+ \). Then, the composite score \( S(f,g) \) is equivalent in probability with the H"older score with \( \gamma > 0 \) and a function \( \phi \).
The proof is found in Appendix C.2.

In the first case of Theorem 4, the integrability of $f \log g$ is assumed for $f, g \in \mathcal{F}$ such that $\mathcal{F}_0 \subset \mathcal{F}$, implying that $\{\omega \in \Omega \mid f(\omega) > 0\} = (0, 1)^d$ holds for $f \in \mathcal{F}$.

Theorem 2 and Theorem 4 imply that the density power score is characterized by the differentiable, separable and affine invariant Bregman score. Indeed, the score of the form (6) includes the differentiable and separable Bregman score, and the affine invariant score of the form (6) is Hölder score. As shown in Theorem 2, the intersection of the differentiable and separable Bregman scores and the Hölder scores is the density power score.

5 Applications of Hölder scores

We use Hölder scores for regression and robust estimation, and investigate the corresponding statistical properties.

5.1 Asymptotically unbiased estimation for regression problems

We use a composite score for the estimation of conditional probabilities or regression functions. Let $x$ and $y$ be the explanatory variable and objective variable, respectively. Suppose that the i.i.d. samples $(x_i, y_i), i = 1, \ldots, n$ are observed from the joint probability density $p(y|x)r(x)$, where $p(y|x)$ is the conditional probability density of $y$ given $x$ and $r(x)$ is the marginal probability density of $x$. Our concern is to estimate $p(y|x)$ from the samples, and the estimation of the marginal probability $r(x)$ is not required.

To estimate $p(y|x)$, let us define a statistical model $\mathcal{M}$, which is a set of conditional probability densities. Suppose that $p(y|x)$ is realized by the model $\mathcal{M}$, i.e., $p(y|x) \in \mathcal{M}$. On each input vector $x$, the discrepancy between $p(y|x)$ and $q(y|x) \in \mathcal{M}$ is measured by $S(p(\cdot|x), q(\cdot|x))$, where $S$ is a composite score. By averaging the composite score with respect to the marginal distribution, we obtain the averaged composite score

$$\bar{S}(p, q|r) := \int S(p(\cdot|x), q(\cdot|x)) r(x) \, dm(x).$$

which is regarded as the loss of the estimate $q(y|x) \in \mathcal{M}$ under the probability density $p(y|x)r(x)$. From the definition of the composite score, the minimum solution of the averaged composite score with respect to $q \in \mathcal{M}$ is attained at $q(y|x) = p(y|x)$.

Let us consider the empirical approximation of $\bar{S}(p, q|r)$ in (7). If $\bar{S}(p, q|r)$ is represented as the expectation with respect to the joint probability $p(y|x)r(x)$, $\bar{S}(p, q|r)$ can be approximated by the empirical mean of the samples, $\{(x_1, y_1), \ldots, (x_n, y_n)\}$. Otherwise, we need an estimate of the conditional probability $p(y|x)$ to obtain an approximation of $\bar{S}(p, q|r)$. Clearly, the later case is not practical, since our purpose is to estimate $p(y|x)$.

Suppose that for any $r(x)$, the averaged composite score $\bar{S}(p, q|r)$ is represented as the expectation for the probability $p(y|x)r(x)$. Then, $S$ is a Bregman score, i.e, $S(p(\cdot|x), q(\cdot|x))$ is expressed as the expectation with respect to $p(\cdot|x)$. If the Bregman score that is equivalent in probability with the Hölder score is used, the affine invariant estimator is obtained for the estimation of the conditional probability. Here, the affine transformation of the objective variable is considered.
Theorem 2 shows that the Bregman score that is equivalent in probability with the Hölder score is of the form (2). The optimum score estimator using (2) is the minimum solution of

\[
\min_{\gamma \in M} \frac{1}{n} \sum_{i=1}^{n} \left( q(\cdot|x_i)^{1+\gamma} \right)^{\kappa/(1+\gamma)} \left( 1 - \frac{1}{\kappa} \frac{q(y_i|x_i)^{\gamma}}{q(\cdot|x_i)^{1+\gamma}} \right),
\]

where $\gamma > 0$ and $\kappa \geq 1$. The composite score \[5\] provides the Fisher consistent estimator of the conditional probability. The estimator with the density power score (resp. $\gamma$-score) is obtained by setting $\kappa = 1 + \gamma$ (resp. $\gamma = 1$). Though a general family of scores including the density power score and $\gamma$-score was proposed by [10], the score [23] is different from the existing family.

The estimator \[8\] is the equivariant estimator under the affine transformation. Provided the data $(x_i, y_i), i = 1, \ldots, n$, let $(\xi(x_i), \sigma^{-1}(y_i - \mu)), i = 1, \ldots, n$ be the transformed data, where $\xi$ is a one-to-one mapping and $\sigma^{-1}(y - \mu)$ is the affine transformation of $y$. When the model $|\det \sigma| q(\sigma y + \mu | \xi(x))$ defined from $q \in M$ is used to the transformed data, the estimator is given by $|\det \sigma| \tilde{q}(\sigma y + \mu | \xi(x))$, where $\tilde{q}(y|x)$ is the estimator obtained by \[8\] based on the original data.

### 5.2 Robust estimation using Hölder scores

The Bregman scores such as the density power scores and $\gamma$-scores are used for robust estimation \[8, 19\]. Let us consider the robustness property of Hölder scores. In robust statistics, the main concern is to develop statistical methods that are not affected by outliers or other small departures from model assumptions.

The robustness of the estimator is quantified by the breakdown point, influence function and so forth \[23\]. Here, the influence function is used to analyze the robustness of the optimum score estimators. Let us introduce the influence functions briefly. Let $p_\theta(x)$ be a probability density on $\mathbb{R}^d$ with a finite dimensional parameter $\theta \in \Theta \subset \mathbb{R}^k$, and $\delta_z(x)$ be the probability density having a point mass at $x = z$. Given the probability density $p_\varepsilon(x) = (1 - \varepsilon)p_\theta(x) + \varepsilon \delta_z(x)$, let $\theta_\varepsilon$ be the minimizer of $\min_{\delta \in \Theta} S(p_\varepsilon, p_\theta)$, where $S$ is a composite score. For $\varepsilon = 0$, the optimal solution is $\theta_0 = \theta$. The parameter $\theta_\varepsilon$ is the optimum score estimator under the contamination $\delta_z$. The influence function of the optimum score estimator against the contamination $\delta_z$ is defined as

\[
\text{IF}(z; \theta, S) = \lim_{\varepsilon \rightarrow +0} \frac{\theta_\varepsilon - \theta}{\varepsilon}.
\]

The influence function $\text{IF}(z; \theta, S)$ provides several measures of the robustness for the optimum score estimator. An example is the gross error sensitivity $\sup_z ||\text{IF}(z; \theta, S)||$, where $\| \cdot \|$ is the Euclidean norm. The estimator that uniformly minimizes the gross error sensitivity over the parameter space is called the most B(ias)-robust estimator. The most B-robust estimator minimizes the worst-case influence of outliers. For the one-dimensional normal distribution, the median estimator is the most B-robust for the estimation of the mean value \[23\]. On the other hand, the estimator satisfying

\[
\lim_{\|z\| \rightarrow \infty} ||\text{IF}(z; \theta, S)|| = 0 \quad \text{for all } \theta \in \Theta
\]

is called the redescending estimator \[23, 27\]. The redescending property is preferable for stable inference, since the influence of extreme outliers tends to zero. Note that the most B-robust estimator is not necessarily the redescending estimator, and vice versa.
It is known that under the normal distribution, the $\gamma$-score has the redescending property, while the density power score does not [19]. In the following theorem, we present the necessary and sufficient condition that the optimum score estimator using the Hölder score has the redescending property for general statistical models.

**Theorem 5.** Suppose that the function $\phi(z)$ in the Hölder score is second order continuously differentiable around $z = 1$. For the statistical model $p_\theta(x), \theta \in \Theta \subset \mathbb{R}^k$, let $s_\theta(x) \in \mathbb{R}^k$ be the score function of the model, i.e., $(s_\theta(x))_i = \frac{\partial}{\partial \theta_i} \log p_\theta(x), i = 1, \ldots, k$. Let us assume the following conditions:

1. The limiting condition $\lim_{\|z\| \to \infty} p_\theta(z) = 0$ holds for all parameter $\theta$.

2. There exists $\gamma > 0$ satisfying the followings:
   
   (a) $p_\theta \in L_{1+\gamma}^+$ holds for all $\theta$.
   
   (b) $\lim_{\|z\| \to \infty} p_\theta(z)^\gamma s_\theta(z) = 0$ holds for all parameter $\theta$.

   (c) Let $I \in \mathbb{R}^{k \times k}$ be the Hessian matrix of $\phi(\langle p_\theta^\gamma \rangle / \langle p_\theta^{1+\gamma} \rangle)$ at $\theta = \theta^* \in \Theta$, i.e.,

   $I_{ij} = \frac{\partial^2}{\partial \theta_i \partial \theta_j} \left\{ \phi\left( \frac{\langle p_\theta p_\theta^\gamma \rangle}{\langle p_\theta^{1+\gamma} \rangle} \right) \right\} \bigg|_{\theta = \theta^*}, \quad (9)$

   for $i, j = 1, \ldots, k$. The Hessian matrix $I$ is invertible at any $\theta^* \in \Theta$.

   (d) For any $\theta^* \in \Theta$, the integral under the measure $m$ and the differential with respect to $\theta$ for the functions $\langle p_\theta^{1+\gamma} \rangle$ and $\langle p_\theta^\gamma p_\theta \rangle$ are interchangeable in the vicinity of $\theta = \theta^*$. In addition, there exists a parameter $\theta$ such that the integral $\langle p_\theta^{1+\gamma} s_\theta \rangle$ is not equal to the zero vector.

Then, the optimum score estimator using Hölder score with $\gamma > 0$ satisfies the redescending property for arbitrary statistical model satisfying the above conditions if and only if $\phi''(1) = -\gamma(1+\gamma)$ holds. All such estimators have the same asymptotic variance.

The proof is deferred to Appendix [19].

The Hölder score that is equivalent in probability with the $\gamma$-score satisfies $\phi''(1) = -\gamma(1+\gamma)$. Hence, for general parametric models, the optimum score estimator using $\gamma$-score has the redescending property. The Hölder scores with $\phi''(1) = -\gamma(1+\gamma)$ include non-Bregman scores, implying that non-Bregman scores can be useful for statistical inference.

The $\gamma$-score is characterized by the following three conditions, i) affine invariance, ii) applicability to regression problems, and iii) redescending property. Indeed, the function $\phi$ in (3) satisfies $\phi''(1) = -\gamma(1+\gamma) + (\kappa - 1)(1+\gamma)$, and $\phi''(1) = -\gamma(1+\gamma)$ holds only for $\kappa = 1$, i.e., the case of $\gamma$-score. A characterization of $\gamma$-score is also presented in [19]. Comparing to the argument in [19], our characterization is more directly connected with the statistical properties of the optimum score estimator.
6 Conclusion

We introduced the Hölder score that is a class of composite scores, and presented its characterization based on the affine invariance of the associated divergence. We studied the relation between the Hölder score and the conventional proper score, i.e., the Bregman score, and derived a class of Bregman scores that is represented as the mixture form of the density power score and γ-score. We also found that the density power score is the intersection of the separable Bregman scores and Hölder scores. Then, we used Hölder scores for statistical inference including regression problems and robust parameter estimation. The Hölder scores that are applicable to regression problems are given by the intersection of Bregman scores and Hölder scores. The Hölder scores outside of the intersection will not produce asymptotically unbiased estimators for the regression problems. In robust parameter estimation, the redescending property was investigated for Hölder score. We proved that the Hölder score satisfying the mild condition on the function φ yields the robust estimator against extreme outliers. In the class of Hölder scores, only the γ-score provides the robust and asymptotically unbiased estimator for regression problems.

As shown in robust estimation in Section 5.2, the Hölder score other than Bregman score can be useful for statistical inference. In this paper, we focused on composite scores of the form (6). An expansion of (6) may provide a wider class of affine invariant composite scores. The final goal on this line is to specify all the affine invariant composite scores, and to reveal its statistical properties. It is also an interesting future work to identify the composite scores inducing equivariant estimators under a data-transformation other than the affine transformation. Another interesting research direction is to investigate the class of equivariant estimators defined from the proper local scores, which depend on the predictive density through its value and the values of its derivatives [14, 17, 29]. The proper local scores provide practical estimators under large dimensional statistical models, since they can be computed without knowledge of the normalizing constant of the probability densities. The invariance of the proper local scores under data-transformation is an important feature to understand the statistical properties of the associated estimators.

A Hölder divergence

proof of Theorem 7. The Hölder score with γ = 0 is the KL score, which is a strictly proper score as shown by many authors. Let us consider Hölder score $S(f, g)$ with $\gamma > 0$ defined on $\mathcal{F} = L_{1+\gamma}^1$. Provided $f \in \mathcal{F}$ and $g^\gamma \in L_{1+1/\gamma}^1$ for $g \in \mathcal{F}$, the Hölder’s inequality leads to

$$\langle fg^\gamma \rangle \leq \langle f^{1+\gamma} \rangle^{1/(1+\gamma)} \langle g^{1+\gamma} \rangle^{\gamma/(1+\gamma)}$$

for all $f, g \in \mathcal{F}$.

The equality holds if and only if $f$ and $g$ are linearly dependent. From the inequality $\phi(z) \geq -z^{1+\gamma}$ for $z \geq 0$, we have

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

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

$$\geq 0.$$

(Hölder’s inequality)
Suppose that $S(p,q) = S(p,p)$ holds for the probability densities $p,q \in \mathcal{P}$. Then, the equality of Hölder’s inequality should hold. Therefore, $p$ and $q$ are linearly dependent, i.e., there exists a constant $c \in \mathbb{R}$ such that $p = cq$ holds. For the probability densities, the constant $c$ should be 1, and we obtain $p = q$. \hfill \Box



B Bregman scores and Hölder scores

proof of Theorem 2 We prove the first case. Suppose that there exists a strictly monotone increasing function $\xi$ such that

$$-G(g) - \int G_g^*(\omega)(f(\omega) - g(\omega))d\omega = -\xi((-\phi((fg^\gamma)/(g^{1+\gamma}))(g^{1+\gamma}))$$

for all $f, g \in \mathcal{F} = L_{1+\gamma}^+$. Here, the expression $-\xi((-\phi((fg^\gamma)/(g^{1+\gamma}))(g^{1+\gamma}))$ is used instead of $\xi((\phi((fg^\gamma)/(g^{1+\gamma}))(g^{1+\gamma}))$ for a simple expression of the potential function. Substituting $f$ into $g$, we have $G(f) = \xi((f^{1+\gamma}))$. For $\delta \in \mathbb{R}$, the function $A(\delta) = (f + \delta h)^{1+\gamma}$ is differentiable at $\delta = 0$ for all $f \in L_{1+\gamma}^+$ and all $h \in L_{1+\gamma}$, and $A'(0) = (1+\gamma)(f^\gamma h)$ holds [18] Chap. 8. In addition, the differentiability of the potential $G(f)$ is assumed. We prove that the function $\xi$ is differentiable on $\mathbb{R}_+^\circ$. Let $a \in \mathbb{R}$ be a real number with a small absolute value, and let us define $g = (1+a)f \in \mathcal{F}$ for a given $f \in \mathcal{F}$. Then, $(1-\varepsilon)f + \varepsilon g = (1+a\varepsilon)f \in \mathcal{F}$ holds for $\varepsilon$ with $|\varepsilon| < \delta$, where $\delta$ is a small positive constant. Let the function $A(\varepsilon)$ be $A(\varepsilon) = G((1-\varepsilon)f + \varepsilon g) = \xi((1+a\varepsilon)^{1+\gamma}(f^{1+\gamma}))$. For all $f \in \mathcal{F}$, $A(\varepsilon)$ is differentiable at $\varepsilon = 0$. This implies that $\xi(z)$ is differentiable for $z > 0$.

We specify the expression of the function $\xi$. The (sub)gradient of $G(g) = \xi((g^{1+\gamma}))$ at $g \in \mathcal{F}$ is given as

$$G_g^*(\omega) = (1+\gamma)\xi'((g^{1+\gamma}))g^\gamma(\omega).$$

Let $x = (g^{1+\gamma})$ and $z = (fg^\gamma)/(g^{1+\gamma})$ for $f, g \in \mathcal{F}$. Then, $(x, z)$ can take any point in $\mathbb{R}_+^\circ \times \mathbb{R}_+^\circ$. The equation (10) is rewritten as

$$\xi(x) + (1+\gamma)\xi'(x)(x-z) = \xi(0).$$

The continuous function $\phi$ satisfies the conditions in Definition 6, i.e., $\phi(1) = -1$ and $\phi(z) \geq -z^{1+\gamma}$ for $z \geq 0$. Hence, there exists a real number $z_0$ such that $0 \leq z_0 < 1$ and $\phi(z_0) = 0$. Substituting $z = z_0$, we obtain the differential equation of $\xi(x)$,

$$\xi(x) + (1+\gamma)(z_0 - 1)x\xi'(x) = \xi(0).$$

The solution is given as

$$\xi(x) = \xi(0) + cx^{1/(1+\gamma)(1-z_0))},$$

where $c$ is a positive constant. For $\kappa = 1/(1-z_0) \geq 1$, we have $G(f) = (f^{1+\gamma})^{\kappa/(1+\gamma)}$ up to an affine transformation with a positive factor. Note that $(f^{1+\gamma})^{\kappa/(1+\gamma)}$ with $\gamma > 0$ and $\kappa \geq 1$ is convex on $\mathcal{F}$ and strictly convex on $\mathcal{P}$.

Let us consider the second case. Suppose that the potential function $G(f) = (f^{1+\gamma})^{\kappa/(1+\gamma)}$ provides a separable Bregman divergence. Then, $\kappa$ should be $1+\gamma$. \hfill \Box
C Affine invariant divergences

Let $\Omega = \mathbb{R}^d$, $\mathcal{B}$ be the Borel set of $\Omega$, and $m : \mathcal{B} \to \mathbb{R}_+$ be the Lebesgue measure on $(\Omega, \mathcal{B})$.

C.1 The functions $U$ and $V$

We show the proof of Theorem 3. Let us consider a necessary condition that the function (6) provides a composite score.

Lemma 1. Under Assumption 7 and Assumption 2, the equality

$$ V(z) = c \int z U'(z) dz, \quad z > 0 $$

holds, where $c \in \mathbb{R}$ is a non-zero constant.

Proof of Lemma 7. Let $A$ and $B$ be disjoint measurable subsets of $(0,1)^d$ such that $A \cup B = (0,1)^d$, and $m(A)$ and $m(B)$ are positive. For $x = (x_1,x_2) \in \mathbb{R}_+^d \times \mathbb{R}_+$, let us define the function class $f_x \in \mathcal{F}_0 \subset \mathcal{F}$ as

$$ f_x(\omega) = \begin{cases} x_1, & \omega \in A, \\ x_2, & \omega \in B, \\ 0, & \text{otherwise}. \end{cases} $$

For $x = (x_1,x_2)$ and $y = (y_1,y_2)$, we have

$$ \langle f_x U(f_y) \rangle = x_1 U(y_1)m(A) + x_2 U(y_2)m(B), \quad \langle V(f_y) \rangle = V(y_1)m(A) + V(y_2)m(B). $$

Since $S$ is a composite score, the inequality

$$ \psi(x_1 U(y_1)m(A) + x_2 U(y_2)m(B), V(y_1)m(A) + V(y_2)m(B)) \geq \psi(x_1 U(x_1)m(A) + x_2 U(x_2)m(B), V(x_1)m(A) + V(x_2)m(B)) $$

holds for $x_1, x_2, y_1, y_2 > 0$. Hence, we have

$$ \frac{\partial}{\partial y_i} \psi(x_1 U(y_1)m(A) + x_2 U(y_2)m(B), V(y_1)m(A) + V(y_2)m(B)) \bigg|_{y=x} = 0, \quad i = 1,2 $$

$$ \iff \psi_1 x_1 U'(x_1) + \psi_2 V'(x_1) = 0, \quad \psi_1 x_2 U'(x_2) + \psi_2 V'(x_2) = 0, $$

for $i = 1,2$, where $\psi_i$ is evaluated at $\langle f_x U(f_x), V(f_x) \rangle \in \mathbb{R}^2$. From Assumption 2(c), the gradient vector of $\psi$ does not vanish. Therefore, the matrix

$$ \begin{pmatrix} x_1 U'(x_1) & V'(x_1) \\ x_2 U'(x_2) & V'(x_2) \end{pmatrix} $$

is not invertible for all $x_1, x_2 > 0$. Thus, the equality

$$ x_1 U'(x_1)V'(x_2) - x_2 U'(x_2)V'(x_1) = 0 $$

17
should hold for all \( x_1, x_2 > 0 \). Since \( U \) is not a constant function on \( \mathbb{R}_+^o \), there exists \( x_2 > 0 \) such that \( U'(x_2) \neq 0 \). Hence, we obtain the equalities,

\[
V'(z) = czU'(z) \quad \text{and} \quad V(z) = c \int zU'(z)dz, \quad z > 0,
\]

with a non-zero constant \( c \).

Below, we present the proof of Theorem 3.

**Proof of Theorem 3.** We assume \( \Omega = \mathbb{R} \). Extension to the multi-dimensional case is straightforward. For a positive real number \( \sigma \), let us consider the affine transformation \( \omega \mapsto \sigma \omega \) for \( \omega \in \mathbb{R} \). This action induces the transformation of the probability density, \( p(\omega) \mapsto p_\sigma(\omega) = \sigma p(\sigma \omega) \). A simple calculation yields that the divergence \( D(p_\sigma, q_\sigma) \) is given as

\[
D(p_\sigma, q_\sigma) = \psi(\langle pU(\sigma q) \rangle, \langle V(\sigma q)/\sigma \rangle) - \psi(\langle pU(\sigma p) \rangle, \langle V(\sigma p)/\sigma \rangle).
\]

Let us define the function set \( \mathcal{V} \) as

\[
\mathcal{V} = \{ v \in L_0 | v(\omega) = 0 \text{ for all } \omega \notin (0, 1), \langle v \rangle = 0, \text{ and } \|v\|_\infty < 1 \}.
\]

Let \( u(\omega) \) be the probability density of the uniform distribution on the interval \((0, 1)\), i.e., \( u(\omega) \) equals 1 on \((0, 1)\) and 0 otherwise. For \( v \in \mathcal{V} \) and \( \varepsilon \) with \( |\varepsilon| < 1 \), the function \( p = u + \varepsilon v \in \mathcal{F}_0 \) is also a probability density. Let \( q(\omega) \) be a probability density in \( \mathcal{F}_0 \). We see that \( D((u + \varepsilon v)_\sigma, q_\sigma) \) is second order differentiable with respect to \( \sigma \) and \( \varepsilon \) in the vicinity of \((\sigma, \varepsilon) = (1, 0)\). This is confirmed by the dominating convergence theorem. Indeed, around \((\sigma, \varepsilon) = (1, 0)\), the functions, \((u + \varepsilon v)U(\sigma q), V(\sigma q)/\sigma, (u + \varepsilon v)U(\sigma(u + \varepsilon v)) \) and \( V(\sigma(u + \varepsilon v))/\sigma \), and those derivatives are all bounded on the interval \((0, 1)\), and they take zero on the outside of the interval \((0, 1)\). The scale function \( h(\sigma) \) is differentiable around \( \sigma = 1 \) because of the differentiability of \( D((u + \varepsilon v)_\sigma, q_\sigma) \) and the equality \( h(\sigma) = D(u + \varepsilon v, q)/D((u + \varepsilon v)_\sigma, q_\sigma) \). The affine invariance of the divergence yields the equality

\[
\frac{\partial}{\partial \sigma} h(\sigma)D((u + \varepsilon v)_\sigma, q_\sigma) = 0 \tag{11}
\]

for all \( v \in \mathcal{V} \) and arbitrary \( \varepsilon \) with \( |\varepsilon| < 1 \). Therefore, we have

\[
\left. \frac{\partial^2}{\partial \varepsilon \partial \sigma} h(\sigma)D((u + \varepsilon v)_\sigma, q_\sigma) \right|_{\sigma = 1, \varepsilon = 0} = 0.
\]

for all \( v \in \mathcal{V} \). The equality above produces

\[
\int_{\Omega} \{c_1 U(q(\omega)) + c_2 U'(q(\omega))q(\omega)\}v(\omega)dm(\omega) = 0,
\]

for all \( v \in \mathcal{V} \), where \( c_1 \) and \( c_2 \) are some constants. Therefore, there exists another constant \( c_3 \) such that the equality

\[
c_1 U(q(\omega)) + c_2 U'(q(\omega))q(\omega) = c_3
\]

should hold for all \( \omega \in (0, 1) \). Here, \( q \) is an arbitrary probability density satisfying the inequality \( 0 < a < q(\omega) < b \) on the support \((0, 1)\). Since \( a \) and \( b \) can take arbitrary positive numbers such that \( 0 < a < b < 1 < b \), the function \( U \) should satisfy the differential equation

\[
c_1 U(z) + c_2 U'(z)z = c_3, \quad z > 0.
\]
Up to a constant factor, the solution is given as \( U(z) = z^\gamma + c \) or \( U(z) = -\log z + c \). From Lemma 1 we conclude that the corresponding \( V \) is \( V(z) = z^{1+\gamma} \) for \( U(z) = z^\gamma + c \), and \( V(z) = z \) for \( U(z) = -\log z + c \) up to a constant factor. Since the equality \( \lim_{z \to 0} V'(z) = V(0) = 0 \) and the existence of \( \lim_{z \to 0} V'(z) \) are assumed in Assumption 2 (b), the real number \( \gamma \) of \( U(z) = z^\gamma + c \) should be positive.

### C.2 The proof of Theorem 4

#### C.2.1 proof of the case 1

Let the functions \( U \) and \( V \) in (6) be \( U(z) = -\log z + c \) and \( V(z) = z \).

**Proof of the case 1 in Theorem 4.** For \( U(z) = -\log z + c \) and \( V(z) = z \), the composite score is given as \( S(f,g) = \psi((-f \log g + cf), (g)) \). For the probability densities \( p, q \in \mathcal{P} \subset \mathcal{F} \), the composite score satisfies the inequality \( \psi(c-\langle p \log q \rangle, 1) \geq \psi(-\langle p \log q \rangle, 1) \). Hence, the function \( \psi(\cdot, 1) \) should be strictly increasing, since \( -\langle p \log q \rangle \geq -\langle p \log p \rangle \) holds for any distinct \( p, q \) in \( \mathcal{P} \). Therefore, \( S(f,g) \) is equivalent in probability with the KL score.

#### C.2.2 proof of the case 2

We prepare some lemmas.

**Lemma 2.** Suppose \( U(z) = z^\gamma + c \) and \( V(z) = z^{1+\gamma} \). Under the assumption in Theorem 4, there exists a function \( \phi : \mathbb{R} \to \mathbb{R} \) and \( s \in \mathbb{R} \) such that the function \( \psi(x, y) \) in (6) is represented as \( \psi(x, y) = \phi((x-c)/y)y^s \) up to a monotone transformation.

**Proof of Lemma 2.** For \( U(z) = z^\gamma + c \) and \( V(z) = z^{1+\gamma} \), we have \( S(p,q) = \psi(\langle pq^\gamma \rangle + c, \langle q^{1+\gamma} \rangle) \) for \( p, q \in \mathcal{P} \). By replacing \( \psi(x + c, y) \) with \( \psi(x, y) \), the composite score on \( \mathcal{P} \) is represented as \( S(p,q) = \psi(\langle pq^\gamma \rangle, \langle q^{1+\gamma} \rangle) \). For \( p, q \in \mathcal{P} \subset L^+_{1+\gamma} \), the integrals \( \langle p^{1+\gamma} \rangle \) and \( \langle pq^\gamma \rangle \) are finite. Let us consider the affine transformation \( \omega \mapsto \sigma \omega \) on \( \Omega = \mathbb{R} \), where \( \sigma > 0 \). In the same way as the derivation of (11) in the proof of Theorem 3 we have

\[
\frac{\partial}{\partial \sigma} h(\sigma) \left\{ \psi(\sigma^\gamma \langle pq^\gamma \rangle, \sigma^{1+\gamma} \langle q^{1+\gamma} \rangle) - \psi(\sigma^\gamma \langle p^{1+\gamma} \rangle, \sigma^\gamma \langle p^{1+\gamma} \rangle) \right\} \bigg|_{\sigma = 1} = 0,
\]

where \( h(\sigma) \) is the scale function. Let us define \( x = \langle pq^\gamma \rangle, y = \langle q^{1+\gamma} \rangle, z = \langle p^{1+\gamma} \rangle \), and \( s = -\frac{d}{d\sigma} \log h(\sigma) \big|_{\sigma = 1} \in \mathbb{R} \). Then, we have

\[-s\psi(x, y) + x\psi_1(x, y) + y\psi_2(x, y) = -s\psi(z, z) + z\psi_1(z, z) + z\psi_2(z, z).\]

Note that \( (x, y, z) \) are independent variables in an open subset of \( \mathbb{R}^3 \). One can prove this fact by using the implicit function theorem. Thus, the left side of the above equation should be a constant for any \( (x, y) \) in an open subset of \( \mathbb{R}^2 \), since the right side is independent of \( (x, y) \). Hence, there exists a real number \( b \in \mathbb{R} \) such that

\[-s\psi(x, y) + x\psi_1(x, y) + y\psi_2(x, y) = b.\]
The general solution of this partial differential equation is found from Euler’s equation [14]. Here, we solve the above PDE by using the variable change. For the polar coordinate system \((r, \theta)\) of \(\mathbb{R}^2\) with \(x = r \cos \theta\) and \(y = r \sin \theta\), the above PDE is expressed as

\[-s \tilde{\psi}(r, \theta) + r \frac{\partial}{\partial r} \tilde{\psi}(r, \theta) = b,\]

where \(\tilde{\psi}(r, \theta) = \psi(r \cos \theta, r \sin \theta)\). All solutions are given by

\[\tilde{\psi}(r, \theta) = \hat{\phi}(\theta) r^s + \begin{cases} -b/s, & s \neq 0, \\ b \log r, & s = 0, \end{cases}\]

where \(\hat{\phi}(\theta)\) is a function of \(\theta\). In the \((x, y)\)-coordinate system, there exists a function \(\phi\) such that

\[\psi(x, y) = \phi(x/y) y^s + \begin{cases} c_1, & s \neq 0, \\ c_0 \log y, & s = 0, \end{cases}\]

where \(c_0, c_1 \in \mathbb{R}\). Without loss of generality we set \(c_1 = 0\). For \(s = 0\), we have \(e^{\psi(x, y)} = e^{\phi(x/y)} y^{c_0}\). Hence \(\psi(x, y)\) or \(e^{\psi(x, y)}\) can be expressed as the form of \(\phi(x/y) y^s\) with \(s \in \mathbb{R}\).

Let \(U(z) = z^\gamma + c\) and \(V(z) = z^{1+\gamma}\) with \(\gamma > 0\) and \(c \in \mathbb{R}\). Then, Lemma 2 ensures that for \(f \in \mathcal{P}\) and \(g \in \mathcal{F}\), the affine invariant composite score is of the form

\[H(f, g) = \phi \left( \frac{\langle fg^\gamma \rangle}{\langle g^{1+\gamma} \rangle} \right) \langle g^{1+\gamma} \rangle^s.\]

with \(s \in \mathbb{R}\) up to a monotone transformation. The sign of the parameter \(s\) is determined by the following lemma.

**Lemma 3.** For \(\gamma > 0\), let \(\mathcal{F} = L_{1+\gamma}^+\) and \(\mathcal{P} = \{p \in \mathcal{F} | \langle p \rangle = 1\}\). Suppose that \(H(f, g)\) in [13] is the composite score on \(\mathcal{P} \times \mathcal{F}\), i.e., \(H(f, g) \geq H(f, f)\) for all \((f, g) \in \mathcal{P} \times \mathcal{F}\), and \(H(p, q) = H(p, p)\) for \((p, q) \in \mathcal{P} \times \mathcal{P}\) implies \(p = q\). Then, \(s > 0 > \phi(1)\) and \(\phi(z) \geq \phi(1) z^{(1+\gamma)s}\) for \(z \geq 0\) hold.

**Proof of Lemma 3.** Remember that the Hölder’s inequality is represented as

\[\langle fg^\gamma \rangle \leq \langle f^{1+\gamma} \rangle^{1/(1+\gamma)} \langle g^{1+\gamma} \rangle^{\gamma/(1+\gamma)}, \quad f, g \in \mathcal{F} = L_{1+\gamma}^+.\]

The equality holds if and only if \(f\) and \(g\) are linearly dependent.

First of all, we prove \(\phi(1) \neq 0\) and \(s \neq 0\). Suppose that \(\phi(1) = 0\) holds. Then, the equality

\[H(p, q) - H(p, p) = \phi \left( \frac{\langle pq^\gamma \rangle}{\langle q^{1+\gamma} \rangle} \right) \langle q^{1+\gamma} \rangle^s = 0\]

holds for \(p, q \in \mathcal{P}\) if and only if \(p = q\). Let \(q\) be the probability density of the uniform distribution on \((0, 1)^d \subset \Omega = \mathbb{R}^d\). Then, arbitrary probability density \(p\) whose support is included in \((0, 1)^d\) satisfies \(H(p, q) - H(p, p) = \phi(1) = 0\). This contradicts the assumption that \(H\) is the composite score. Therefore, \(\phi(1) \neq 0\) holds. Suppose \(s = 0\). Then, the equality

\[H(p, q) - H(p, p) = \phi \left( \frac{\langle pq^\gamma \rangle}{\langle q^{1+\gamma} \rangle} \right) - \phi(1) = 0\]
holds for $p, q \in P$ if and only if $p = q$. In the same way as above, setting $q$ as the probability density of the uniform distribution on $(0, 1)^d$ yields the contradiction. Therefore, we obtain $s \neq 0$.

Next, we prove $\phi(0) \geq 0 > \phi(1)$. Let $A$ and $B$ be disjoint subsets of $\Omega = \mathbb{R}^d$, and suppose that they have finite positive measures. Let $p$ and $q$ be the probability densities of the uniform distribution on $A$ and $B$, respectively. Then, we have $(p^{1+\gamma}) = m(A)^{-\gamma}, (q^{1+\gamma}) = m(B)^{-\gamma}$ and $(pq^\gamma) = 0$. For the composite score $H(p,q)$, the inequality

$$H(p,q) - H(p,p) = \phi(0)m(B)^{-\gamma s} - \phi(1)m(A)^{-\gamma s} \geq 0$$

holds. For $\gamma > 0$ and $s \neq 0$, $m(A)^{-\gamma s}$ and $m(B)^{-\gamma s}$ can take any positive real numbers independently. Hence, the inequality $\phi(0) \geq 0 > \phi(1)$ should hold. This result and $\phi(1) \neq 0$ lead to $\phi(0) \geq 0 > \phi(1)$.

Let us consider the sign of $s$. Since $H$ is the composite score, the inequality

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

holds for all $f \in P$ and $g \in F$. There exist $f \in P$ and $g \in F$ such that

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

(15)

holds, i.e., the Hölder’s inequality strictly holds with $1 = \langle fg^\gamma \rangle/\langle g^{1+\gamma} \rangle$. For example, for linearly independent functions, $f \in P$ and $g_0 \in F$, 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 $f \in P, g \in F$ satisfying (15), we have the inequality

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

from the non-negativity of $H(f, g) - H(f, f)$ and positivity of $(f^{1+\gamma})$. From $0 < \langle g^{1+\gamma} \rangle/\langle f^{1+\gamma} \rangle < 1$, $\phi(1) < 0$ and $s \neq 0$, the inequality above holds only when $s > 0$.

Suppose that there exists $z_0 > 0$ such that $\phi(z_0) < \phi(1)z_0^{(1+\gamma)s}$ holds. Choose $f \in P$ and $g \in F$ 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, $g = f/z_0 \in F$ for some $f \in P$. For such $f$ and $g$, we have

$$H(f, g) - H(f, f) = \phi(z_0)\langle g^{1+\gamma} \rangle^s - \phi(1)\langle f^{1+\gamma} \rangle^s$$

$$< \phi(1)z_0^{(1+\gamma)s} \langle g^{1+\gamma} \rangle^s - \phi(1)\langle f^{1+\gamma} \rangle^s$$

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

$$= 0,$$

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 $z > 0$. From $\phi(0) \geq 0$ and $(1 + \gamma)s > 0$, eventually the inequality $\phi(z) \geq \phi(1)z^{(1+\gamma)s}$ should hold for all $z \geq 0$. \qed
Finally, we prove the case 2 of Theorem 4.

Proof of the case 2 in Theorem 4. From Lemma 2 and Lemma 3 the affine invariant composite score is expressed as

$$H(p, q) = \phi \left( \frac{\langle p \gamma \rangle}{\langle q^{1+\gamma} \rangle} \right) \langle q^{1+\gamma} \rangle^s \quad \text{for} \quad p, q \in P,$$

with $\gamma > 0$, where $\phi(z) \geq \phi(1)z^{(1+\gamma)s}$ for $z \geq 0$ and $s > 0 > \phi(1)$ hold. The transformation using the strictly increasing function $\xi(H) = |H/\phi(1)|^{1/s}\text{sign}(H)$ ensures that the composite score $H$ is equivalent in probability with the H"older score with $\phi$.

Theorem 4.

Proof of Theorem 4. Let us define $p_z = (1-\varepsilon)p_{\theta^*} + \varepsilon \delta_z(x) = p_{\theta^*} + \varepsilon(\delta_z(x) - p_{\theta^*}(x))$, and $r_z(x) = r_z(x) = \delta_z(x) - p_{\theta^*}(x)$. By using the implicit function theorem to the $R^k$-valued function

$$\langle \theta, \varepsilon \rangle \mapsto \frac{\partial}{\partial \theta} \left\{ \phi \left( \frac{\langle p_{\theta^*} \gamma \rangle}{\langle p_{\theta^*}^{1+\gamma} \rangle} \right) \langle r_z \gamma \rangle \right\}$$

around $(\theta, \varepsilon) = (\theta^*, 0)$, we obtain

$$\text{IF}(z, \theta^*, S) = -I^{-1} \frac{\partial}{\partial \theta} \left\{ \phi' \left( \frac{\langle p_{\theta^*} \gamma \rangle}{\langle p_{\theta^*}^{1+\gamma} \rangle} \right) \langle r_z \gamma \rangle \right\} \bigg|_{\theta = \theta^*}. \quad (16)$$

Hence, the estimator has the redescending property if and only if

$$\lim_{\|z\| \to \infty} \frac{\partial}{\partial \theta} \left\{ \phi' \left( \frac{\langle p_{\theta^*} \gamma \rangle}{\langle p_{\theta^*}^{1+\gamma} \rangle} \right) \langle r_z \gamma \rangle \right\} \bigg|_{\theta = \theta^*} = 0$$

holds for any $\theta^* \in \Theta$. From the assumption on $\phi$, we have $\phi'(1) = -1 - \gamma$. A calculation using $\phi(1) = -1$ and $\phi'(1) = -1 - \gamma$ yields that the derivative in the above is given as

$$\frac{\partial}{\partial \theta} \phi' \left( \frac{\langle p_{\theta^*} \gamma \rangle}{\langle p_{\theta^*}^{1+\gamma} \rangle} \right) \langle r_z \gamma \rangle \bigg|_{\theta = \theta^*} = -\phi''(1) \frac{\langle r_z \gamma \rangle}{\langle p_{\theta^*} \gamma \rangle} \int p_{\theta^*}(x)^{1+\gamma}s_{\theta^*}(x)dm(x)$$

$$= \gamma(1 + \gamma) \int r_z(x)p_{\theta^*}(x)^{1+\gamma}s_{\theta^*}(x)dm(x),$$

in which the interchangeability of the integral and differential is used. From the assumption, the limiting of $\|z\| \to \infty$ leads to

$$\lim_{\|z\| \to \infty} \frac{\partial}{\partial \theta} \left\{ \phi' \left( \frac{\langle p_{\theta^*} \gamma \rangle}{\langle p_{\theta^*}^{1+\gamma} \rangle} \right) \langle r_z \gamma \rangle \right\} \bigg|_{\theta = \theta^*}$$

$$= \phi''(1) \int p_{\theta^*}(x)^{1+\gamma}s_{\theta^*}(x)dm(x).$$
The expression above vanishes for all $\theta^*$ if and only if the equality $\phi''(1) = -\gamma(1 + \gamma)$ holds.

The asymptotic variance of the estimator is determined from the influence function. Some calculation shows that Hölder score affects the influence function via $\phi''(1)$. Hence, the optimum score estimators using Hölder scores with the same $\phi''(1)$ have the same asymptotic variance. □

References

[1] J. D. Abernethy and R. M. Frongillo. A characterization of scoring rules for linear properties. *Journal of Machine Learning Research - Proceedings Track*, 23:27.1–27.13, 2012.

[2] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. Clustering with Bregman divergences. *J. Mach. Learn. Res.*, 6:1705–1749, December 2005.

[3] A. Basu, I. R. Harris, N. L. Hjort, and M. C. Jones. Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85(3):549–559, 1998.

[4] A. Basu, H. Shioya, and C. Park. *Statistical Inference: The Minimum Distance Approach*. Monographs on Statistics and Applied Probability. Taylor & Francis, 2010.

[5] J. O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer Series in Statistics. Springer, 1985.

[6] J. M. Borwein and Q. Q. J. Zhu. *Techniques of variational analysis*. CMS books in mathematics. Springer Science+Business Media, Incorporated, 2005.

[7] L. M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex prog ramming. *USSR Computational Mathematics and Mathematical Physics*, 7:200–217, 1967.

[8] B. J. Bremnes. Probabilistic forecasts of precipitation in terms of quantiles using nwp model output. *Monthly Weather Review*, 132:338–347, 2004.

[9] G. W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Rev.*, 78:1–3, 1950.

[10] A. Cichocki and S. Amari. Families of alpha- beta- and gamma- divergences: Flexible and robust measures of similarities. *Entropy*, 12(6):1532–1568, 2010.

[11] M. Collins, R. E. Schapire, and Y. Singer. Logistic regression, adaboost and Bregman distances. In *Proceedings of the Thirteenth Annual Conference on Computational Learning Theory*, pages 158–169, 2000.

[12] A. P. Dawid. Coherent measures of discrepancy, uncertainty and dependence, with applications to bayesian predictive experimental design. Technical report, University College London, Dept. of Statistical Science, 1998.

[13] A. P. Dawid. The geometry of proper scoring rules. *Annals of the Institute of Statistical Mathematics*, 59(1):77–93, 2007.

[14] A. P. Dawid, S. Lauritzen, and M. Parry. Proper local scoring rules on discrete sample spaces. *Annals of Statistics*, 40:593–608, 2012.
[15] D. Duffie and J. Pan. An overview of value at risk. *Journal of Derivatives*, 4:749, 1997.

[16] S. Eguchi, O. Komori, and S. Kato. Projective power entropy and maximum Tsallis entropy distributions. *Entropy*, 13:1746–1764, 2011.

[17] W. Ehm and T. Gneiting. Local proper scoring rules of order two. *Annals of Statistics*, 40:609–637, 2012.

[18] M. Fabian, P. Habala, P. Hájek, V. Montesinos Santalucia, J. Pelant, and V. Zizler. *Functional Analysis and Infinite-Dimensional Geometry*. CMS Books in Mathematics. Springer, 2001.

[19] H. Fujisawa and S. Eguchi. Robust parameter estimation with a small bias against heavy contamination. *J. Multivar. Anal.*, 99(9):2053–2081, 2008.

[20] T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102:359–378, 2007.

[21] I. J. Good. Comment on ”measuring information and uncertainty,” by R. J. Buehler. In V. P. Godambe and D. A. Sprott, editors, *Foundations of Statistical Inference*, page 337339, Toronto: Holt, Rinehart and Winston, 1971.

[22] P. D. Grünwald and A. P. Dawid. Game theory, maximum entropy, minimum discrepancy, and robust bayesian decision theory. *Annals of Statistics*, 32:1367–1433, 2003.

[23] F. R. Hampel, P. J. Rousseeuw, E. M. Ronchetti, and W. A. Stahel. *Robust Statistics. The Approach based on Influence Functions*. John Wiley and Sons, Inc., 1986.

[24] A. D. Hendrickson and R. J. Buehler. Proper scores for probability forecasters. *The Annals of Mathematical Statistics*, 42:19161921, 1971.

[25] P. J. Huber. Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35(1):73–101, 1964.

[26] M. C. Jones, N. L. Hjort, I. R. Harris, and A. Basu. A comparison of related density-based minimum divergence estimators. *Biometrika*, 88(3):865–873, 2001.

[27] R. Maronna, R.D. Martin, and V. Yohai. *Robust Statistics: Theory and Methods*. Wiley, 2006.

[28] N. Murata, T. Takenouchi, T. Kanamori, and S. Eguchi. Information geometry of U-Boost and Bregman divergence. *Neural Computation*, 16(7):1437–1481, 2004.

[29] M. Parry, A. P. Dawid, and S. Lauritzen. Proper local scoring rules. *Annals of Statistics*, 40:561–592, 2012.

[30] C. Tsallis. Possible generalization of Boltzmann-Gibbs statistics. *Journal of Statistical Physics*, 52(1-2):479–487, 1988.

[31] K. Tsuda, G. Rätsch, and M. K. Warmuth. Matrix exponentiated gradient updates for on-line learning and Bregman projection. *J. Mach. Learn. Res.*, 6:995–1018, 2005.

[32] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2000.