Bernoulli **20**(4), 2014, 2278–2304
DOI: 10.3150/13-BEJ557

Affine invariant divergences associated with proper composite scoring rules and their applications

TAKAFUMI KANAMORI[1] and HIRONORI FUJISAWA[2]

[1]*Department of Computer Science and Mathematical Informatics, Nagoya University, Furocho Chikusaku, Nagoya 464-8601, Japan. E-mail: kanamori@is.nagoya-u.ac.jp*
[2]*The Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan. E-mail: fujisawa@ism.ac.jp*

In statistical analysis, measuring a score of predictive performance is an important task. In many scientific fields, appropriate scoring rules were tailored to tackle the problems at hand. A proper scoring rule is a popular tool to obtain statistically consistent forecasts. Furthermore, a mathematical characterization of the proper scoring rule was studied. As a result, it was revealed that the proper scoring rule 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 scoring rules as an extension of the typical scoring rules in order to obtain a wider class of probabilistic forecasting. Then, we propose a class of composite scoring rules, 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 no 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 scoring rules for statistical forecasting.

Keywords: affine invariance; Bregman score; composite scoring rule; divergence; Hölder score

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, for example, weather and climate prediction [8,9], computational finance [16], 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 scoring rule is formalized as a function taking two inputs, that is, a probability distribution for the prediction and an outcome. In order to achieve high prediction performance on average, ideally, optimization of the expected scor-

1350-7265 © 2014 ISI/BS

ing rule is conducted. When the identically and independently distributed (i.i.d.) samples are available, the expected scoring rule is approximated by the empirical mean over the samples. By optimizing the empirical mean 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 scoring rules [9,14,21,25]. We regard the scoring rule as a loss to be minimized. The estimator obtained from the scoring rule is called the optimum score estimator. To obtain a good estimator, scoring rules need to satisfy some assumptions. A typical assumption is that the scoring rule is proper. Given a probability distribution of outcomes, the optimal value of the expected proper scoring rule is attained by setting the prediction probability to be the true probability distribution. Under mild assumptions, optimization of the proper scoring rule averaged over the observed samples produces a statistically consistent estimator. The proper scoring rule is a special case of M-estimation [26], and the statistical properties of the proper scoring rule have been studied in the framework of M-estimators [33], Chapter 5.

The proper scoring rule is a basic element that yields important concepts in statistical inference. According to [14], the proper scoring rule 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. Bregman divergence [7] is an important class of divergences, since it is closely related to the proper scoring rule. A major milestone in the theoretical approach is the characterization of the proper scoring rule by using the Bregman divergence [1,21,25]. More precisely, any proper scoring rule produces a Bregman divergence, and a given Bregman divergence yields a proper scoring rule. The correspondence established a way to investigate the proper scoring rule by using the Bregman divergence on statistical models.

In the present paper, we introduce composite scoring rules as an extension of the proper scoring rules in order to obtain a wider class of probabilistic forecasting. Then, we propose a class of composite scoring rules, 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, that is, $\omega \mapsto \sigma^{-1}(\omega - \mu)$ for the data $\omega \in \mathbb{R}^d$, where σ is a d by d invertible matrix and μ 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 θ 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 $\widehat{\theta}$ of the mean value θ should be transformed into $\sigma^{-1}(\widehat{\theta} - \mu)$. 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 scoring rules and the Bregman divergences, the composite scoring rules correspond to a class of divergences. When the divergence is invariant under the data transformation, the corresponding composite scoring rule provides an equivariant estimator. We prove that the Hölder scores are characterized by the affine invariance of the associated divergence, that is, among a class of composite scoring rules, only Hölder scores provide 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], scoring rules of continuous variables have so far received little attention. In this paper, our main concern is the scoring rules 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 scoring rules and associated divergences. Bregman scores and their separable variant are also introduced as an important class of composite scoring rules. Then, we show a way to use composite scoring rules 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 scoring rules.

2. Composite scoring rules and divergences

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

Let us summarize the notations to be used throughout the paper. Let $\mathbb{R}$ be the set of all real numbers. The nonnegative 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°. Thus, $\mathbb{R}_+^\circ$ implies the set of all positive real numbers, that is, $\mathbb{R}_+^\circ = \{x \in \mathbb{R} \mid x > 0\}$. For a sample space Ω, let $\mathcal{B}$ be a σ-algebra of subsets of Ω and $m : \mathcal{B} \to \mathbb{R}_+$ be a σ-finite measure on $(\Omega, \mathcal{B})$. When Ω is a subset of a Euclidean space, $\mathcal{B}$ and m denote the Borel algebra and Lebesgue measure, respectively. The set of all measurable functions on Ω is denoted as L_0, that is, $L_0 = \{f : \Omega \to \mathbb{R} \mid f \text{ is measurable on } (\Omega, \mathcal{B}, m)\}$. For $f \in L_0$, the integral $\int_\Omega f(\omega)\,\mathrm{d}m(\omega)$ is denoted as $\langle f \rangle$. Let $\|\cdot\|_\alpha$ for $1 \leq \alpha < \infty$ be the L_α-norm, that is, $\|f\|_\alpha = \langle |f|^\alpha \rangle^{1/\alpha}$, and $\|\cdot\|_\infty$ be the essential sup-norm. For $\alpha \geq 1$, let L_α be $L_\alpha = \{f \in L_0 \mid \|f\|_\alpha < \infty\}$. For $\alpha = 0$ or $\alpha \geq 1$, L_α^+ denotes the set of all nonnegative and nonzero functions in L_α, that is, $L_\alpha^+ = \{f \in L_\alpha \mid f \geq 0, f \neq 0\}$. Provided a set of measurable and nonnegative functions $\mathcal{F} \subset L_0^+$, $\mathcal{P}$ denotes the set of probability densities in $\mathcal{F}$, that is, $\mathcal{P} = \{p \in \mathcal{F} \mid \langle p \rangle = 1\}$, in which the dependency on $\mathcal{F}$ is dropped if there is no confusion. For a differentiable function ψ, ψ_i with the integer i denotes the partial derivative of ψ with respect to the ith argument, for example, for $\psi(x, y)$, ψ_1 and ψ_2 denote $\frac{\partial \psi}{\partial x}$ and $\frac{\partial \psi}{\partial y}$, 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 scoring rule of the forecast using q. When the probability density of the outcome is p, the expected scoring rule is given as

$$S_0(p, q) := \int_\Omega S_0(\omega, q) p(\omega)\, \mathrm{d}m(\omega).$$

Suppose that the expected scoring rule 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.

We expand the expected scoring rules to more general forms. It is defined not only for probability densities but also nonnegative functions.

Definition 2.1 (Composite scoring rule; proper composite scoring rule; entropy). *Let $\mathcal{F}$ be a convex subset in L_0^+, and the set of probability densities in $\mathcal{F}$ is denoted as $\mathcal{P}$, that is, $\mathcal{P} = \{p \in \mathcal{F} \mid \langle p \rangle = 1\}$. The function $S(f, g) : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$ is called the* composite scoring rule *on $\mathcal{F}$, if $S(f, g)$ is of the form*

$$S(f, g) = T\left(\int_\Omega S_0(\omega, g) f(\omega)\, \mathrm{d}m(\omega), g\right), \tag{2.1}$$

where $S_0 : \Omega \times \mathcal{F} \to \mathbb{R}$ and $T : \mathbb{R} \times \mathcal{F} \to \mathbb{R}$. The function $S_0(\cdot, g) f(\cdot)$ is assumed to be integrable for all $f, g \in \mathcal{F}$. The composite scoring rule $S(f, g)$ satisfying the following two conditions is called proper composite scoring rule:

1. *$S(f, g) \geq S(f, f)$ for all $f, g \in \mathcal{F}$.*
2. *For $p, q \in \mathcal{P}$, $S(p, q) = S(p, p)$ implies $p = q$ (almost surely).*

The function $S(f, f)$ is referred to as entropy.

When the composite scoring rule $S(f, g)$ is defined only on the set of probability densities and the function T is given as $T(c, g) = c$, the composite scoring rule in the above is reduced to the expected scoring rule. Likewise, the proper composite scoring rule is an extension of the expected strictly proper scoring rule [14,21,23,25], that is regarded as the proper composite scoring rule with $T(c, g) = c$. In Section 3, we propose a class of proper composite scoring rules with a nontrivial T.

Remark 2.1. In our definition, the domain of the composite scoring rule is not necessarily a set of probability densities, but it can be a set of nonnegative functions. In [25], the strictly proper scoring rules are characterized on the set of nonnegative functions. The definition in the present paper simplifies mathematical analysis on composite scoring rules.

Let us define the equivalence class on composite scoring rules. The meaning of the equivalence class is illustrated in Section 2.3.

Definition 2.2 (Equivalence of composite scoring rules). *The composite scoring rules $S(f,g)$ and $\widetilde{S}(f,g)$ on $\mathcal{F}$ are equivalent if there exists a strictly increasing function $\xi:\mathbb{R}\to\mathbb{R}$ such that $\widetilde{S}(f,g)=\xi(S(f,g))$ holds for all $f,g\in\mathcal{F}$. The composite scoring rules $S(f,g)$ and $\widetilde{S}(f,g)$ on $\mathcal{F}$ are equivalent in probability if there exists a strictly increasing function $\xi:\mathbb{R}\to\mathbb{R}$ such that $\widetilde{S}(p,q)=\xi(S(p,q))$ holds for all probability densities $p,q\in\mathcal{P}\subset\mathcal{F}$.*

A different definition of the equivalence class was also proposed by [13,14], in which the classical expected scoring rules $S(p,q)$ and $\widetilde{S}(p,q)$ on $\mathcal{P}$ are equivalent, if there exist a positive constant $c>0$ and a function $k:\mathcal{P}\to\mathbb{R}$ such that $\widetilde{S}(p,q)=cS(p,q)+k(p)$ holds. The equivalence class in Definition 2.2 is more suitable for our analysis.

Definition 2.3 (Divergence). *Let S be a proper composite scoring rule on $\mathcal{F}$. Then, we call*

$$D(f,g)=S(f,g)-S(f,f),\qquad f,g\in\mathcal{F},$$

the divergence associated with S.

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

2.2. Bregman scores

As an important class of proper composite scoring rules, we introduce Bregman scores and their separable variant. Under a mild assumption, any strictly proper scoring rule on $\mathcal{P}$ is expressed as a Bregman score on $\mathcal{P}$; see [1,21,25] for details.

Definition 2.4 (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*

$$G(f)\ge G(g)+\int_\Omega G_g^*(\omega)f(\omega)\,\mathrm{d}m(\omega)-\int_\Omega G_g^*(\omega)g(\omega)\,\mathrm{d}m(\omega),\qquad \text{for } f,g\in\mathcal{F}$$

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

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

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

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], Chapter 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 scoring rule (2.1) with

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

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 is a proper composite scoring rule. 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 scoring rule; see Theorem 1 of [21].

The separable variant of the Bregman score is defined below.

Definition 2.5 (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, where $J(f)$ implies the composition function of J and $f \in \mathcal{F}$, that is, $J(f)(\omega) = J(f(\omega))$. The separable Bregman divergence is the divergence associated with the separable Bregman score.*

The separable Bregman score is of the form

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

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

We show some examples of Bregman scores and associated divergences.

Example 2.1 (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

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

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. The entropy $S(p, p)$ of $p \in \mathcal{P}$ is given as $\langle -p \log p \rangle + 1$, that is equal to the differential entropy [12] up to a constant.

Example 2.2 (Density power score). Let $\mathcal{F}$ be $\mathcal{F} = L_{1+\gamma}^+$ for a given $\gamma > 0$. The density power score on $\mathcal{F}$ is defined as

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

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 entropy $S(p, p)$ of $p \in \mathcal{P}$ is given as $-\langle p^{1+\gamma} \rangle / \gamma$, that is, the Tsallis entropy $(1 - \langle p^{1+\gamma} \rangle)/\gamma$ [31] up to a constant. The integrability of $f g^\gamma$ is confirmed by Hölder's inequality. The associated divergence is called the density power divergence [3,4,27]. When the parameter γ in the density power divergence tends to zero, the KL-divergence is recovered.

Example 2.3 (γ-score; pseudospherical score). Let $\mathcal{F}$ be $\mathcal{F} = L^+_{1+\gamma}$ for a given $\gamma > 0$. The pseudospherical score [22] is defined as

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

which is a nonseparable 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 scoring rule, $-\frac{1}{\gamma}\log(-S(f,g))$, is called the γ-score in this paper. The γ-score is proposed in [17,20], and it is used for robust parameter estimation. As the limiting case of $\gamma \to 0$, the divergence associated with the γ-score recovers KL-divergence.

2.3. Optimum score estimator

Statistical inference using the composite scoring rule (2.1) is conducted by substituting the empirical probability and the model probability into the composite scoring rule. Provided the i.i.d. samples $\omega_1, \ldots, \omega_n$ from the probability density p, the empirical approximation of $S(p,q)$ for a given probability density q is given as

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

where $\widetilde{p}$ denotes the empirical probability. For a sufficiently large number of samples, $S(\widetilde{p}, q)$ 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 $S(\widetilde{p}, q)$ with respect to q over a statistical model. The estimator $\widehat{q}$ is called the *optimum score estimator* [21]. The estimator using the strictly proper scoring rule is a special case of M-estimation [26], and its statistical properties have been deeply investigated [33].

When two proper composite scoring rules are equivalent in probability in the sense of Definition 2.2, they produce the same estimator.

3. Hölder scores

In this section, we propose a class of composite scoring rules, named Hölder scores, a part of which is not represented as the Bregman score. The relation between the Hölder scores and Bregman scores is also presented.

3.1. Definition of Hölder score

Bregman scores are widely used for statistical inference [2,11,29,32], 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 γ-score are used for robust estimation [3,20]. In this section, we propose a class of composite scoring rules called *Hölder scores*. 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.

Definition 3.1 (Hölder score). *The Hölder score with a nonnegative parameter γ 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*

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

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

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

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

We prove the basic property of the Hölder score.

Theorem 3.1. *The Hölder score is a proper composite scoring rule.*

The proof of Theorem 3.1 is found in Appendix A. Theorem 3.1 ensures that the Hölder score leads to the associated divergence, that is referred to as the Hölder divergence.

The Hölder score with $\gamma > 0$ is represented as the composite scoring rule (2.1) 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 nonnegativity of the Hölder divergence. The entropy of the Hölder score is $S(f, f) = -\langle f^{1+\gamma}\rangle$, which is in agreement with the Tsallis entropy [31] up to an affine transformation.

An appropriate choice of the function ϕ produces the composite scoring rule equivalent with the density power score or γ-score. Indeed, the Hölder score with the lower bound $\phi(z) = -z^{1+\gamma}$ yields $S(f, g) = -\langle fg^{\gamma}\rangle^{1+\gamma}/\langle g^{1+\gamma}\rangle^{\gamma}$ that is equivalent with γ-score. The density power score is equivalent with the Hölder score with $\phi(z) = \gamma - (1 + \gamma)z$.

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 3.2 (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 called differentiable. The corresponding Bregman score (resp. divergence) is called the differentiable Bregman score (resp. divergence).

The differentiability above makes our analysis rather simple. For nondifferentiable 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 nonlinear optimization techniques are directly applicable to obtain the optimum score estimator.

Theorem 3.2. *Let the function ϕ in the Hölder score be 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.

Remember that the KL score is a differentiable and separable Bregman. Hence, the intersection of (separable) Bregman score and Hölder score is equivalent with 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

$$S(f, g) = \langle g^{1+\gamma} \rangle^{\kappa/(1+\gamma)} \left(1 - \frac{1}{\kappa} - \frac{\langle f g^{\gamma} \rangle}{\langle g^{1+\gamma} \rangle} \right). \tag{3.1}$$

The above Bregman scores include the density power score ($\kappa = 1 + \gamma$) and γ-score ($\kappa = 1$) in each equivalent class. The Hölder score corresponding to the Bregman score (3.1) is given by the function $\phi(z)$ defined as

$$\phi(z) = -\kappa^{(1+\gamma)/\kappa} |z - 1 + 1/\kappa|^{(1+\gamma)/\kappa} \operatorname{sign}(z - 1 + 1/\kappa), \tag{3.2}$$

where $\operatorname{sign}(z)$ 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 proper composite scoring rule (3.1).

4. Affine invariance of Hölder divergence

Affine transformation of the observed data is often used in statistical analysis. Let $\Omega = \mathbb{R}^d$, $\mathcal{B}$ be the Borel set of Ω, and m be the Lebesgue measure on $(\Omega, \mathcal{B})$. 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 $\omega_1, \dots, \omega_n \in \mathbb{R}^d$, let the vector μ be the sample mean of the observations, and the matrix σ

be the diagonal matrix such that the kth diagonal element is equal to the sample-based standard deviation of the kth component of the observed data. Then, each element of the transformed data, $\sigma^{-1}(\omega_1 - \mu), \dots, \sigma^{-1}(\omega_n - \mu)$, 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 $\widehat{q}$ be the estimator of p based on the original data $\{\omega_1, \dots, \omega_n\}$, and $\widehat{q_{\sigma,\mu}}$ be the estimator based on the transformed data, $\{\sigma^{-1}(\omega_1 - \mu), \dots, \sigma^{-1}(\omega_n - \mu)\}$. It will be natural to require that the estimator is transformed in a consistent way, when the data is transformed, that is, the equality

$$(\widehat{q})_{\sigma,\mu} = \widehat{q_{\sigma,\mu}} \tag{4.1}$$

should hold. The estimators enjoying (4.1) do not essentially depend on the choice of the units in the measurement. In the present paper, the estimator satisfying (4.1) is called the *affine invariant estimator*. In a formal mathematical description, the term invariant decision rule or 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 scoring rules satisfying the equality $S(p, q) = S(p_{\sigma,\mu}, q_{\sigma,\mu})$. However, the equality is not necessity. In the below, we introduce proper composite scoring rules and associated divergences that provide affine invariant estimators.

Definition 4.1 (Affine invariant divergence; affine invariant proper composite scoring rule). *Let S be a proper composite scoring rule on $\mathcal{F}$, and D be the associated divergence. The divergence D is affine invariant if there exists an $\mathbb{R}_+^\circ$-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*

$$h(\sigma, \mu) D(p_{\sigma,\mu}, q_{\sigma,\mu}) = D(p, q) \tag{4.2}$$

holds for any pair of probability densities $p, q \in \mathcal{P}$ and arbitrary affine transformation. The function h is called the scale function. The proper composite scoring rule S inducing the affine invariant divergence is called the affine invariant proper composite scoring rule.

Remark 4.1. The affine invariance of the divergence is equivalent with the condition that there exist functions $h(\sigma, \mu)$ and $k(\sigma, \mu, p)$ such that $h(\sigma, \mu) S(p_{\sigma,\mu}, q_{\sigma,\mu}) + k(\sigma, \mu, p) = S(p, q)$ holds for the proper composite scoring rule S. Indeed, (4.2) leads to $h(\sigma, \mu) S(p_{\sigma,\mu}, q_{\sigma,\mu}) - h(\sigma, \mu) S(p_{\sigma,\mu}, p_{\sigma,\mu}) + S(p, p) = S(p, q)$. The affine invariance of the proper composite scoring rule is described by using the function k.

We briefly prove that the affine invariant proper composite scoring rule provides the affine invariant estimator. Let S be an affine invariant proper composite scoring rule, and $\widehat{q}$ 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,

$$D(p,\widehat{q}) \le D(p,q) \quad \text{and} \quad D\big(p_{\sigma,\mu},(\widehat{q})_{\sigma,\mu}\big) \le D(p_{\sigma,\mu},q_{\sigma,\mu})$$

hold for all $q\in\mathcal{M}$. On the other hand, $\widehat{q_{\sigma,\mu}}$ is the minimum solution of $\min_{q_{\sigma,\mu}} D(p_{\sigma,\mu},q_{\sigma,\mu})$, when the model $\{q_{\sigma,\mu} \mid q\in\mathcal{M}\}$ is used. Therefore, the equivariant property (4.1) holds, if the optimal solution is unique.

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

$$\begin{aligned} D(p_{\sigma,\mu},q_{\sigma,\mu}) &= \phi\left(\frac{\langle p_{\sigma,\mu}q_{\sigma,\mu}^{\gamma}\rangle}{\langle q_{\sigma,\mu}^{1+\gamma}\rangle}\right)\langle q_{\sigma,\mu}^{1+\gamma}\rangle + \langle p_{\sigma,\mu}^{1+\gamma}\rangle \\ &= \phi\left(\frac{|\det\sigma|^{\gamma}\langle pq^{\gamma}\rangle}{|\det\sigma|^{\gamma}\langle q^{1+\gamma}\rangle}\right)\langle q^{1+\gamma}\rangle|\det\sigma|^{\gamma} + \langle p^{1+\gamma}\rangle|\det\sigma|^{\gamma} \\ &= |\det\sigma|^{\gamma} D(p,q). \end{aligned}$$

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

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

that is, $\mathcal{F}_0\subset\mathcal{F}\subset L_0^+$ holds.

The subset $(0,1)^d$ in the above assumption can be replaced with any subset with a finite measure.

For proper composite scoring rules, we assume the following conditions.

Assumption 4.2 (Assumption on the proper composite scoring rule). *For the proper composite scoring rule $S(f,g)$, we assume three conditions*:

(a) *$S(f,g)$ has the form of*

$$S(f,g)=\psi\big(\langle fU(g)\rangle,\langle V(g)\rangle\big) \qquad \text{for all } f,g\in\mathcal{F}, \tag{4.3}$$

where U and V are real-valued functions on $\mathbb{R}_+$ and ψ is a function on a subset of $\mathbb{R}^2$, that is, S is the proper composite scoring rule (2.1) with $S_0(\omega, g) = U(g(\omega))$ and $T(c, g) = \psi(c, \langle V(g)\rangle)$. For all $f, g \in \mathcal{F}$, the functions $f(\omega)U(g(\omega))$ and $V(g(\omega))$ are integrable.

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

(c) *Let $D_{U,V}$ and $E_{U,V}$ be subsets of $\mathbb{R}^2$ defined as*

$$D_{U,V} = \{(\langle fU(g)\rangle, \langle V(g)\rangle) \in \mathbb{R}^2 \mid f, g \in \mathcal{F}\},$$
$$E_{U,V} = \{(\langle fU(f)\rangle, \langle V(f)\rangle) \in \mathbb{R}^2 \mid f \in \mathcal{F}\},$$

respectively. For arbitrary point $x \in D_{U,V}$, there exists an open neighbourhood of x on which ψ 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 (ψ_1, ψ_2) does not vanish.

Let us explain the assumption (4.3). In general, the composite scoring rule $S(f, g)$ is described as a functional of g. Hence, the assumption (4.3) is a strong restriction on the class of composite scoring rules. On the other hand, all separable Bregman scores and Hölder scores are expressed by the form of (4.3). In addition, the proper composite scoring rules of the form (4.3) are not included in the class of Bregman scores. In practice, the composite scoring rule (4.3) is tractable, since it can be calculated via integrals.

In Assumption 4.2(b), we assumed $V(0) = 0$ in order to guarantee the integrability of the function whose support is not equal to Ω. More precisely, let $Z = \{\omega \in \mathbb{R}^d \mid f(\omega) = 0\}$ with $m(Z) = \infty$, then $\langle V(g)\rangle = \int_Z V(0)\,\mathrm{d}m + \int_{\Omega\setminus Z} V(g)\,\mathrm{d}m$ will not be finite unless $V(0) = 0$. In Assumption 4.2(c), we assumed that the gradient vector (ψ_1, ψ_2) does not become the zero vector at $(\langle fU(f)\rangle, \langle V(f)\rangle)$. 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 4.2(c).

The functions U and V of the affine invariant proper composite scoring rules are determined by Theorem 4.1.

Theorem 4.1. *Let S be an affine invariant proper composite scoring rule. Suppose that Assumption 4.1 and Assumption 4.2 hold. Then, the functions U and V in (4.3) 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 C.1. For each possibility of U and V, the composite scoring rule is identified in the following theorem.

Theorem 4.2. *Let S be an affine invariant proper composite scoring rule. Suppose that Assumption 4.1 and Assumption 4.2 hold.*

1. *Let us define $U(z) = -\log z + c$ and $V(z) = z$ in (4.3). Then, $S(f, g)$ 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 (4.3), and let $\mathcal{F}$ be $\mathcal{F} = L^+_{1+\gamma}$. Then, $S(f, g)$ is equivalent in probability with the Hölder score with $\gamma > 0$ and a function ϕ.*

The proof is found in Appendix C.2.

In the first case of Theorem 4.2, 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 all $f \in \mathcal{F}$.

Theorem 3.2 and Theorem 4.2 lead to the fact that the density power score is characterized by the differentiable, separable and affine invariant Bregman score. Let us consider the equivalence class of proper composite scoring rules induced by Definition 2.2. The proper composite scoring rule of the form (4.3) includes the differentiable and separable Bregman score, and the affine invariant proper composite scoring rule of the form (4.3) is Hölder score. Therefore, as shown in Theorem 3.2, the intersection of the differentiable and separable Bregman scores and the Hölder scores is given by 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 composite scoring rules 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), $i = 1, \dots, n$ 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 $\mathcal{M}$, that is a set of conditional probability densities. Suppose that $p(y|x)$ is realized by the model $\mathcal{M}$, that is, $p(y|x) \in \mathcal{M}$. 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 proper composite scoring rule. By averaging $S(p(\cdot|x), q(\cdot|x))$ with respect to the marginal distribution, we obtain the averaged scoring rule

$$\bar{S}(p, q|r) := \int S\big(p(\cdot|x), q(\cdot|x)\big) r(x)\, \mathrm{d}m(x) \tag{5.1}$$

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 proper composite scoring rule, the minimum solution of the averaged scoring rule with respect to $q \in \mathcal{M}$ is attained at $q(y|x) = p(y|x)$.

Let us consider the empirical approximation of $\bar{S}(p, q|r)$. 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. 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 scoring rule $\bar{S}(p,q|r)$ is represented as the expectation for the probability $p(y|x)r(x)$. Then, S is a Bregman score, that is, $S(p(\cdot|x),q(\cdot|x))$ 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 3.2 shows that the Bregman score that is equivalent in probability with the Hölder score is of the form (3.1). Given samples $(x_1,y_1),\dots,(x_n,y_n)$, the optimum score estimator using (3.1) is the minimum solution of

$$\min_{q\in\mathcal{M}} \frac{1}{n}\sum_{i=1}^{n}\langle q(\cdot|x_i)^{1+\gamma}\rangle^{\kappa/(1+\gamma)}\left(1-\frac{1}{\kappa}-\frac{q(y_i|x_i)^{\gamma}}{\langle q(\cdot|x_i)^{1+\gamma}\rangle}\right), \tag{5.2}$$

where $\gamma>0$, $\kappa\geq 1$, and $\langle q(\cdot|x)^{1+\gamma}\rangle=\int_\Omega q(y|x)^{1+\gamma}\,\mathrm{d}m(y)$. The optimization problem (5.2) provides the Fisher consistent estimator of the conditional probability. The estimator with the density power score (resp. γ-score) is obtained by setting $\kappa=1+\gamma$ (resp. $\gamma=1$). A family of scoring rules including the density power score and γ-score was proposed by [10]. The proper composite scoring rule (3.1) is different from the existing one.

The estimator (5.2) is the equivariant estimator under the affine transformation. Provided the data (x_i,y_i), $i=1,\dots,n$, let $(\xi(x_i),\sigma^{-1}(y_i-\mu))$, $i=1,\dots,n$ be the transformed data, where ξ is a one-to-one mapping and $\sigma^{-1}(y-\mu)$ is the affine transformation of y. When the model $|\det\sigma|q(\sigma y+\mu\mid\xi(x))$ defined from $q\in\mathcal{M}$ is used to the transformed data, the estimator is given by $|\det\sigma|\widehat{q}(\sigma y+\mu\mid\xi(x))$, where $\widehat{q}(y|x)$ is the estimator obtained by (5.2) based on the original data.

5.2. Robust estimation using Hölder scores

The Bregman scores such as the density power scores and γ-scores are used for robust estimation [3,20]. 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 [24]. 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 shifted Dirac's delta function having the point mass at $x=z$, that is conventionally regarded as a probability density. Given the probability density $p_\varepsilon(x)=(1-\varepsilon)p_\theta(x)+\varepsilon\delta_z(x)$, let θ_ε be the minimizer of $\min_{\bar{\theta}\in\Theta}S(p_\varepsilon,p_{\bar{\theta}})$, where S is a proper composite scoring rule. For $\varepsilon=0$, the optimal solution is $\theta_0=\theta$. The parameter θ_ε is the optimum score estimator under the contamination δ_z. The influence function of the optimum score estimator against the contamination δ_z is defined as

$$\mathrm{IF}(z;\theta,S)=\lim_{\varepsilon\to+0}\frac{\theta_\varepsilon-\theta}{\varepsilon}.$$

The influence function $\mathrm{IF}(z;\theta,S)$ provides several measures of the robustness for the optimum score estimator.

An example is the gross error sensitivity $\sup_z \|\mathrm{IF}(z;\theta,S)\|$, 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 [24]. Let $p_\theta(x)$ be the probability density of the one-dimensional normal distribution with the mean $\theta\in\mathbb{R}$ and the unit variance. For the model $p_\theta(x)$, the scoring rule producing the median estimator is given by the expectation of $S_0(x,p_\theta)=(-2\log(\sqrt{2\pi}p_\theta(x)))^{1/2}=|x-\theta|$. For general statistical models, however, this scoring rule will not be proper, and may not be well-defined.

In this section, our main concern is another robustness measure called *redescending property*. The estimator satisfying

$$\lim_{\|z\|\to\infty}\big\|\mathrm{IF}(z;\theta,S)\big\|=0 \qquad \text{for all } \theta\in\Theta$$

is called the redescending estimator [24,28]. 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 γ-score has the redescending property, while the density power score does not [20]. 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 5.1. *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 Fisher's efficient score function of the model, that is, $(s_\theta(x))_i=\frac{\partial}{\partial\theta_i}\log p_\theta(x)$, $i=1,\ldots,k$. Let us assume the following conditions*:

1. *The limiting condition $\lim_{\|z\|\to\infty}p_\theta(z)=0$ holds for all parameter θ.*
2. *There exists $\gamma>0$ satisfying the followings*:
   (a) *$p_\theta\in L_{1+\gamma}^{+}$ holds for all θ.*
   (b) *$\lim_{\|z\|\to\infty}p_\theta(z)^\gamma s_\theta(z)=0$ holds for all parameter θ.*
   (c) *Let $I\in\mathbb{R}^{k\times k}$ be the Hessian matrix of $\phi(\langle p_{\theta^*}p_\theta^\gamma\rangle/\langle p_\theta^{1+\gamma}\rangle)\langle p_\theta^{1+\gamma}\rangle$ at $\theta=\theta^*\in\Theta$, that is,*
   $$I_{ij}=\frac{\partial^2}{\partial\theta_i\,\partial\theta_j}\left\{\phi\left(\frac{\langle p_{\theta^*}p_\theta^\gamma\rangle}{\langle p_\theta^{1+\gamma}\rangle}\right)\langle p_\theta^{1+\gamma}\rangle\right\}\bigg|_{\theta=\theta^*}, \tag{5.3}$$
   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 θ for the functions $\langle p_\theta^{1+\gamma}\rangle$ and $\langle p_{\theta^*}p_\theta^\gamma\rangle$ are interchangeable in the vicinity of $\theta=\theta^*$. In addition, there exists a parameter θ 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 D.

The Hölder score equivalent in probability with the γ-score satisfies $\phi''(1) = -\gamma(1+\gamma)$. Hence, for general parametric models, the optimum score estimator using γ-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.

Based on the above argument, the γ-score is characterized by the following three conditions, (i) affine invariance, (ii) applicability to regression problems, and (iii) redescending property. Indeed, the function ϕ in (3.2) satisfies $\phi''(1) = -\gamma(1+\gamma) + (\kappa - 1)(1+\gamma)$, and $\phi''(1) = -\gamma(1+\gamma)$ holds only for $\kappa = 1$, that is, the case of γ-score. A characterization of γ-score is also presented in [20]. Comparing to the argument in [20], our characterization is more directly connected with the statistical properties of the optimum score estimator.

6. Conclusion

We introduced the Hölder scores, a class of proper composite scoring rules, and presented their characterization based on the affine invariance of the associated divergences. We showed the relation between the Hölder scores and the conventional proper scoring rules, that is, the Bregman scores, and derived a class of Bregman scores (3.1) that is represented as the mixture form of the density power score and γ-score. We also proved that the intersection of the separable Bregman scores and Hölder scores is equivalent with the density power score. Furthermore, we applied the Hölder scores to statistical inference including regression problems and robust parameter estimation. Among the equivalent class of composite scoring rules, the Hölder scores 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 proper composite scoring rules of the form (4.3). An expansion of (4.3) may provide a wider class of affine invariant proper composite scoring rules. The final goal on this line is to specify all the affine invariant proper composite scoring rules, and to reveal its statistical properties. It is also an interesting future work to identify the composite scoring rules 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 scoring rules, which depend on the predictive density through its value and the values of its derivatives [15,18,30]. The proper local scoring rules 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 scoring rules under data-transformations is an important feature to understand the statistical properties of the associated estimators.

Appendix A: Hölder divergence

Proof of Theorem 3.1. The Hölder score with $\gamma = 0$ 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 $\mathcal{F} = L^+_{1+\gamma}$. Provided $f \in \mathcal{F}$ and $g^\gamma \in L^+_{1+1/\gamma}$ for $g \in \mathcal{F}$, the Hölder's inequality leads to

$$\langle fg^\gamma \rangle \le \langle f^{1+\gamma} \rangle^{1/(1+\gamma)} \langle g^{1+\gamma} \rangle^{\gamma/(1+\gamma)} \qquad \text{for all } f, g \in \mathcal{F}.$$

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

$$\begin{aligned} S(f,g) - S(f,f) &= \phi\left(\frac{\langle fg^\gamma \rangle}{\langle g^{1+\gamma} \rangle}\right)\langle g^{1+\gamma} \rangle + \langle f^{1+\gamma} \rangle \\ &\ge -\left(\frac{\langle fg^\gamma \rangle}{\langle g^{1+\gamma} \rangle}\right)^{1+\gamma}\langle g^{1+\gamma} \rangle + \langle f^{1+\gamma} \rangle \\ &\ge 0 \qquad \text{(Hölder's inequality)}. \end{aligned}$$

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, that is, 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 $p = q$. □

Appendix B: Bregman scores and Hölder scores

Proof of Theorem 3.2. We prove the first case. Suppose that there exists a strictly monotone increasing function ξ such that

$$-G(g) - \int G^*_g(\omega)\big(f(\omega) - g(\omega)\big)\,\mathrm{d}m(\omega) = -\xi\left(-\phi\big(\langle fg^\gamma \rangle / \langle g^{1+\gamma} \rangle\big)\langle g^{1+\gamma} \rangle\right) \tag{B.1}$$

for all $f, g \in \mathcal{F} = L^+_{1+\gamma}$. Here, the expression $-\xi(-\phi(\langle fg^\gamma \rangle / \langle g^{1+\gamma} \rangle)\langle g^{1+\gamma} \rangle)$ is used instead of $\xi(\phi(\langle fg^\gamma \rangle / \langle g^{1+\gamma} \rangle)\langle g^{1+\gamma} \rangle)$ for a simple expression of the potential function. Substituting f into g, we have $G(f) = \xi(\langle f^{1+\gamma} \rangle)$. For $\delta \in \mathbb{R}$, the function $A(\delta) = \langle |f + \delta h|^{1+\gamma} \rangle$ is differentiable at $\delta = 0$ for all $f \in L^+_{1+\gamma}$ and all $h \in L_{1+\gamma}$, and $A'(0) = (1 + \gamma)\langle f^\gamma h \rangle$ holds [19], Chapter 8. In addition, the differentiability of the potential $G(f)$ is assumed. We prove that the function ξ 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 ε with

$|\varepsilon| < \delta$, where δ is a small positive constant. Let the function $A(\varepsilon)$ be $A(\varepsilon) = G((1-\varepsilon)f + \varepsilon g) = \xi((1+a\varepsilon)^{1+\gamma}\langle f^{1+\gamma}\rangle)$. 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 ξ. The (sub)gradient of $G(g) = \xi(\langle g^{1+\gamma}\rangle)$ at $g \in \mathcal{F}$ is given as

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

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

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

The continuous function ϕ satisfies the conditions in Definition 3.1, that is, $\phi(1) = -1$ 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 $G(f) = \langle f^{1+\gamma}\rangle^{\kappa/(1+\gamma)}$ up to an affine transformation with a positive factor. Note that $\langle f^{1+\gamma}\rangle^{\kappa/(1+\gamma)}$ 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 $G(f) = \langle f^{1+\gamma}\rangle^{\kappa/(1+\gamma)}$ provides a separable Bregman divergence. Then, κ should be $1+\gamma$. □

Appendix C: Affine invariant divergences

Let $\Omega = \mathbb{R}^d$, $\mathcal{B}$ be the Borel set of Ω, 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 4.1. Let us consider a necessary condition that the function (4.3) provides a proper composite scoring rule.

Lemma C.1. *Under Assumption 4.1 and Assumption 4.2, the equality*

$$V(z) = c\int zU'(z)\,\mathrm{d}z, \qquad z > 0$$

holds, where $c \in \mathbb{R}$ is a nonzero constant.

Proof. 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 $x = (x_1, x_2) \in \mathbb{R}_+^\circ \times \mathbb{R}_+^\circ$, 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

$$\begin{aligned} \langle f_x U(f_y) \rangle &= x_1 U(y_1) m(A) + x_2 U(y_2) m(B), \\ \langle V(f_y) \rangle &= V(y_1) m(A) + V(y_2) m(B). \end{aligned}$$

Since S is a proper composite scoring rule, the inequality

$$\begin{aligned} &\psi\big(x_1 U(y_1) m(A) + x_2 U(y_2) m(B), V(y_1) m(A) + V(y_2) m(B)\big) \\ &\quad \geq \psi\big(x_1 U(x_1) m(A) + x_2 U(x_2) m(B), V(x_1) m(A) + V(x_2) m(B)\big) \end{aligned}$$

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

$$\begin{aligned} &\frac{\partial}{\partial y_i} \psi\big(x_1 U(y_1) m(A) + x_2 U(y_2) m(B), V(y_1) m(A) + V(y_2) m(B)\big)\Big|_{y=x} = 0 \\ &\iff \quad \psi_1 x_1 U'(x_1) + \psi_2 V'(x_1) = 0, \qquad \psi_1 x_2 U'(x_2) + \psi_2 V'(x_2) = 0, \end{aligned}$$

for $i = 1, 2$, where ψ_i is evaluated at $(\langle f_x U(f_x) \rangle, \langle V(f_x) \rangle) \in \mathbb{R}^2$. From Assumption 4.2(c), the gradient vector of ψ 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

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

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

$$V'(z) = c z U'(z) \quad \text{and} \quad V(z) = c \int z U'(z)\, \mathrm{d}z, \qquad z > 0,$$

with a nonzero constant c. □

Below, we present the proof of Theorem 4.1.

Proof of Theorem 4.1. We assume $\Omega = \mathbb{R}$. Extension to the multi-dimensional case is straightforward. For a positive real number σ, 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

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

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

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

Let $u(\omega)$ be the probability density of the uniform distribution on the interval (0, 1), that is, $u(\omega)$ equals 1 on the interval (0, 1) and 0 otherwise. For $v \in \mathcal{V}$ and ε with $|\varepsilon| < 1$, the function $p = u + \varepsilon v \in \mathcal{F}_0$ is also a probability density, where $\mathcal{F}_0$ is the function set defined in Assumption 4.1. Let $q(\omega)$ be a probability density in $\mathcal{F}_0$. We see that $D((u+\varepsilon v)_\sigma, q_\sigma)$ is second order continuously differentiable with respect to σ and ε 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

$$\frac{\partial}{\partial\sigma} h(\sigma) D\big((u+\varepsilon v)_\sigma, q_\sigma\big) = 0 \tag{C.1}$$

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

$$\frac{\partial^2}{\partial\varepsilon\,\partial\sigma} h(\sigma) D\big((u+\varepsilon v)_\sigma, q_\sigma\big)\Big|_{\substack{\sigma=1\\ \varepsilon=0}} = 0$$

for all $v \in \mathcal{V}$. The equality above with some algebra produces

$$\int_\Omega \big\{c_1 U\big(q(\omega)\big) + c_2 U'\big(q(\omega)\big)q(\omega)\big\} v(\omega)\,\mathrm{d}m(\omega) = 0,$$

for all $v \in \mathcal{V}$, where c_1 and c_2 are constants. In the above calculation, equalities such as $\langle vU'(u)\rangle = 0$ or $\langle vV'(u)\rangle = 0$ are used with the convention $0 \times (\pm\infty) = 0$ in Lebesgue integration. Therefore, there exists another constant c_3 such that the equality

$$c_1 U\big(q(\omega)\big) + c_2 U'\big(q(\omega)\big)q(\omega) = c_3$$

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 $0 < a < 1 < b$, the function U should satisfy the differential equation

$$c_1 U(z) + c_2 U'(z) z = c_3, \qquad z > 0.$$

Up to a constant factor, the solution is given as $U(z) = z^\gamma + c$ or $U(z) = -\log z + c$. From Lemma C.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 $\lim_{z\searrow 0} V(z)=V(0)=0$ and the existence of $\lim_{z\searrow 0} V'(z)$ are assumed in Assumption 4.2(b), the real number γ of $U(z)=z^\gamma+c$ should be positive. □

C.2. The proof of Theorem 4.2

C.2.1. *Proof of the case 1*

Let the functions U and V in (4.3) be $U(z)=-\log z+c$ and $V(z)=z$.

Proof of the case 1 in Theorem 4.2. For $U(z)=-\log z+c$ and $V(z)=z$, the proper composite scoring rule is given as $S(f,g)=\psi(\langle -f\log g+cf\rangle,\langle g\rangle)$. For the probability densities $p,q\in\mathcal{P}\subset\mathcal{F}$, the proper composite scoring rule satisfies the inequality $\psi(c-\langle p\log q\rangle,1)\geq\psi(c-\langle p\log p\rangle,1)$. Hence, the function $\psi(\cdot,1)$ should be strictly increasing, since $-\langle p\log q\rangle>-\langle p\log p\rangle$ 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 C.2. *Suppose $U(z)=z^\gamma+c$ and $V(z)=z^{1+\gamma}$. Under the assumption in Theorem 4.2, there exists a function $\phi:\mathbb{R}\to\mathbb{R}$ and $s\in\mathbb{R}$ such that the function $\psi(x,y)$ in (4.3) is represented as $\psi(x,y)=\phi((x-c)/y)y^s$ up to a monotone transformation.*

Proof. For $U(z)=z^\gamma+c$, $V(z)=z^{1+\gamma}$, 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 proper composite scoring rule 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 (C.1) in the proof of Theorem 4.1, we have

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

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 $s=-\frac{\mathrm{d}}{\mathrm{d}\sigma}\log h(\sigma)|_{\sigma=1}\in\mathbb{R}$. 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 [15]. Here, we solve the above PDE by using the variable change. For the polar coordinate system (r,θ) of $\mathbb{R}^2$ with $x=r\cos\theta$ and $y=r\sin\theta$, the above PDE is expressed as

$$-s\bar{\psi}(r,\theta)+r\frac{\partial}{\partial r}\bar{\psi}(r,\theta)=b, \tag{C.2}$$

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

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

where $\bar{\phi}(\theta)$ is a function of θ. In the (x,y)-coordinate system, there exists a function ϕ 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 $\mathrm{e}^{\psi(x,y)}=\mathrm{e}^{\phi(x/y)}y^{c_0}$. Hence, $\psi(x,y)$ or $\mathrm{e}^{\psi(x,y)}$ 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 C.2 ensures that for $f\in\mathcal{P}$ and $g\in\mathcal{F}$, the affine invariant proper composite scoring rule is of the form

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

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

Lemma C.3. *For $\gamma>0$, let $\mathcal{F}=L^+_{1+\gamma}$ and $\mathcal{P}=\{p\in\mathcal{F}\mid\langle p\rangle=1\}$. Suppose that $H(f,g)$ in (C.3) is the proper composite scoring rule on $\mathcal{P}\times\mathcal{F}$, that is, $H(f,g)\geq H(f,f)$ 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. Remember that the Hölder's inequality is represented as

$$\langle fg^\gamma\rangle\leq\langle f^{1+\gamma}\rangle^{1/(1+\gamma)}\langle g^{1+\gamma}\rangle^{\gamma/(1+\gamma)},\qquad f,g\in\mathcal{F}=L^+_{1+\gamma}. \tag{C.4}$$

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 proper composite scoring rule. 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 $p,q\in\mathcal{P}$ 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 $\langle p^{1+\gamma}\rangle=m(A)^{-\gamma}$, $\langle q^{1+\gamma}\rangle=m(B)^{-\gamma}$ and $\langle pq^{\gamma}\rangle=0$. For the proper composite scoring rule $H(p,q)$, 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 $\phi(0)\geq 0\geq\phi(1)$ 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 proper composite scoring rule, 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 $f\in\mathcal{P}$ and $g\in\mathcal{F}$. There exist $f\in\mathcal{P}$ and $g\in\mathcal{F}$ such that

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

holds, that is, the Hölder's inequality strictly holds with $1=\langle fg^{\gamma}\rangle/\langle g^{1+\gamma}\rangle$. For example, for linearly independent functions, $f\in\mathcal{P}$ and $g_0\in\mathcal{F}$, 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 $f\in\mathcal{P}$, $g\in\mathcal{F}$ satisfying (C.5), 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 nonnegativity of $H(f,g)-H(f,f)$ and positivity of $\langle f^{1+\gamma}\rangle$. 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 $f\in\mathcal{P}$ and $g\in\mathcal{F}$ 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, $g = f/z_0 \in \mathcal{F}$ for some $f \in \mathcal{P}$. For such f and g, we have

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

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 $z > 0$. 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 $z \geq 0$. □

Finally, we prove the case 2 of Theorem 4.2.

Proof of the case 2 in Theorem 4.2. From Lemma C.2 and Lemma C.3, the affine invariant proper composite scoring rule is expressed as

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

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 $\xi(H) = |H/\phi(1)|^{1/s}\,\mathrm{sign}(H)$ ensures that the proper composite scoring rule H is equivalent in probability with the Hölder score with $\gamma > 0$. The inequality $\phi(z) \geq \phi(1)z^{(1+\gamma)s}$ with $\phi(1) < 0$ is transformed into $\phi(z) \geq -z^{1+\gamma}$. □

Appendix D: Redescending property

For a differentiable real-valued function $f(\theta)$ of $\theta \in \mathbb{R}^k$, let $\frac{\partial f}{\partial \theta}$ be the gradient column vector of $f(\theta)$.

Proof of Theorem 5.1. Let us define $p_\varepsilon = (1-\varepsilon)p_{\theta^*} + \varepsilon\delta_z(x) = p_{\theta^*} + \varepsilon(\delta_z(x) - p_{\theta^*}(x))$, and $r_z(x)$ be $r_z(x) = \delta_z(x) - p_{\theta^*}(x)$. By using the implicit function theorem to the $\mathbb{R}^k$-valued function

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

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

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

See [24], Section 4.2, for details. Hence, the estimator has the redescending property if and only if

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

holds for any $\theta^*\in\Theta$. From the assumption on ϕ, 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

$$\begin{aligned}\frac{\partial}{\partial\theta}\phi'\left(\frac{\langle p_{\theta^*}p_\theta^\gamma\rangle}{\langle p_\theta^{1+\gamma}\rangle}\right)\langle r_z p_\theta^\gamma\rangle\bigg|_{\theta=\theta^*} &= -\phi''(1)\frac{\langle r_z p_{\theta^*}^\gamma\rangle}{\langle p_{\theta^*}^{1+\gamma}\rangle}\int p_{\theta^*}(x)^{1+\gamma}s_{\theta^*}(x)\,\mathrm{d}m(x)\\ &\quad-\gamma(1+\gamma)\int r_z(x)p_{\theta^*}(x)^\gamma s_{\theta^*}(x)\,\mathrm{d}m(x),\end{aligned}$$

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

$$\begin{aligned}&\lim_{\|z\|\to\infty}\frac{\partial}{\partial\theta}\left\{\phi'\left(\frac{\langle p_{\theta^*}p_\theta^\gamma\rangle}{\langle p_\theta^{1+\gamma}\rangle}\right)\langle r_z p_\theta^\gamma\rangle\right\}\bigg|_{\theta=\theta^*}\\ &=\big(\phi''(1)+\gamma(1+\gamma)\big)\int p_{\theta^*}(x)^{1+\gamma}s_{\theta^*}(x)\,\mathrm{d}m(x).\end{aligned}$$

The expression above vanishes for all θ^* 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. □

Acknowledgements

The authors are grateful to the anonymous reviewers for their helpful comments. TK was partially supported by JSPS KAKENHI Grant Number 24500340.

References

[1] Abernethy, J.D. and Frongillo, R.M. (2012). A characterization of scoring rules for linear properties. *J. Mach. Learn. Res.*: *Workshop and Conference Proceedings* **23** 27.1–27.13.

[2] Banerjee, A., Merugu, S., Dhillon, I.S. and Ghosh, J. (2005). Clustering with Bregman divergences. *J. Mach. Learn. Res.* **6** 1705–1749. MR2249870

[3] Basu, A., Harris, I.R., Hjort, N.L. and Jones, M.C. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika* **85** 549–559. MR1665873

[4] Basu, A., Shioya, H. and Park, C. (2011). *Statistical Inference*: *The Minimum Distance Approach*. *Monographs on Statistics and Applied Probability* **120**. Boca Raton, FL: CRC Press. MR2830561

[5] Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. *Springer Series in Statistics*. New York: Springer. MR0804611

[6] Borwein, J.M. and Zhu, Q.J. (2005). *Techniques of Variational Analysis*. *CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC* **20**. New York: Springer. MR2144010

[7] Brègman, L.M. (1967). A relaxation method of finding a common point of convex sets and its application to the solution of problems in convex programming. *Ž. Vyčisl. Mat. i Mat. Fiz.* **7** 620–631. MR0215617

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

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

[10] Cichocki, A. and Amari, S.-i. (2010). Families of alpha- beta- and gamma-divergences: Flexible and robust measures of similarities. *Entropy* **12** 1532–1568. MR2659408

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

[12] Cover, T.M. and Thomas, J.A. (2006). *Elements of Information Theory*, 2nd ed. Hoboken, NJ: Wiley. MR2239987

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

[14] Dawid, A.P. (2007). The geometry of proper scoring rules. *Ann. Inst. Statist. Math.* **59** 77–93. MR2396033

[15] Dawid, A.P., Lauritzen, S. and Parry, M. (2012). Proper local scoring rules on discrete sample spaces. *Ann. Statist.* **40** 593–608. MR3014318

[16] Duffie, D. and Pan, J. (1997). An overview of value at risk. *J. Derivatives* **4** 7–49.

[17] Eguchi, S., Komori, O. and Kato, S. (2011). Projective power entropy and maximum Tsallis entropy distributions. *Entropy* **13** 1746–1764. MR2851127

[18] Ehm, W. and Gneiting, T. (2012). Local proper scoring rules of order two. *Ann. Statist.* **40** 609–637. MR3014319

[19] Fabian, M., Habala, P., Hájek, P., Montesinos Santalucía, V., Pelant, J. and Zizler, V. (2001). *Functional Analysis and Infinite-Dimensional Geometry*. *CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC* **8**. New York: Springer. MR1831176

[20] Fujisawa, H. and Eguchi, S. (2008). Robust parameter estimation with a small bias against heavy contamination. *J. Multivariate Anal.* **99** 2053–2081. MR2466551

[21] Gneiting, T. and Raftery, A.E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* **102** 359–378. MR2345548

[22] Good, I.J. (1971). Comment on "Measuring information and uncertainty," by R.J. Buehler. In *Foundations of Statistical Inference* (V.P. Godambe and D.A. Sprott, eds.) 337–339. Toronto: Holt, Rinehart and Winston.

[23] Grünwald, P.D. and Dawid, A.P. (2004). Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. *Ann. Statist.* **32** 1367–1433. MR2089128

[24] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. (1986). *Robust Statistics*: *The Approach Based on Influence Functions*. *Wiley Series in Probability and Mathematical Statistics*: *Probability and Mathematical Statistics*. New York: Wiley. MR0829458

[25] Hendrickson, A.D. and Buehler, R.J. (1971). Proper scores for probability forecasters. *Ann. Math. Statist.* **42** 1916–1921. MR0314430

[26] Huber, P.J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35** 73–101. MR0161415

[27] Jones, M.C., Hjort, N.L., Harris, I.R. and Basu, A. (2001). A comparison of related density-based minimum divergence estimators. *Biometrika* **88** 865–873. MR1859416

[28] Maronna, R.A., Martin, R.D. and Yohai, V.J. (2006). *Robust Statistics: Theory and Methods. Wiley Series in Probability and Statistics*. Chichester: Wiley. MR2238141

[29] Murata, N., Takenouchi, T., Kanamori, T. and Eguchi, S. (2004). Information geometry of U-boost and Bregman divergence. *Neural Comput.* **16** 1437–1481.

[30] Parry, M., Dawid, A.P. and Lauritzen, S. (2012). Proper local scoring rules. *Ann. Statist.* **40** 561–592. MR3014317

[31] Tsallis, C. (1988). Possible generalization of Boltzmann–Gibbs statistics. *J. Stat. Phys.* **52** 479–487. MR0968597

[32] Tsuda, K., Rätsch, G. and Warmuth, M.K. (2005). Matrix exponentiated gradient updates for on-line learning and Bregman projection. *J. Mach. Learn. Res.* **6** 995–1018. MR2249846

[33] van der Vaart, A.W. (1998). *Asymptotic Statistics. Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge: Cambridge Univ. Press. MR1652247

Received May 2013