Bernoulli **24**(1), 2018, 53–79
DOI: 10.3150/16-BEJ857

Proper scoring rules and Bregman divergence

EVGENI Y. OVCHAROV

Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Acad. Georgi Bonchev Str., Block 8, 1113 Sofia, Bulgaria. E-mail: trulr6@yahoo.com

Proper scoring rules measure the quality of probabilistic forecasts. They induce dissimilarity measures of probability distributions known as Bregman divergences. We survey the literature on both entities and present their mathematical properties in a unified theoretical framework. Score and Bregman divergences are developed as a single concept. We formalize the proper affine scoring rules and present a motivating example from robust estimation. And lastly, we develop the elements of the regularity theory of entropy functions and describe under what conditions a general convex function may be identified as the entropy function of a proper scoring rule and whether this association is unique.

Keywords: Bregman divergence; characterisation; convex analysis; entropy; proper scoring rule; robust estimation; subgradient

1. Introduction

Proper scoring rules measure the quality of probabilistic forecasts. They can be used as protocols for eliciting private information and as incentives for a forecaster to make truthful predictions. Bregman divergences, on the other hand, originate in convex programming as generalizations of the Euclidean metric. The two notions are intimately connected through the fact that they both derive from convex functions known as entropies. The principal objective of the present work is to present and review the mathematical properties of the two entities in a unified theoretical framework. To that end, we systematically generalize a number of results from the Euclidean setting to the framework of spaces of finite signed measures. In this setting score and Bregman divergences may be identified. We formalize the proper affine scoring rules and motivate their interest in the context of estimation problems, which complements the approach of [16,17], who formalize them as truthful mechanisms in decision theory. And finally, we develop the elements of the regularity theory of entropy functions. This allows us to describe under what conditions a general convex function can be identified with the entropy function of a proper scoring rule and whether this association is unique.

1.1. Detailed plan of the paper

In Section 2, we introduce the general mathematical framework of the paper. We also illustrate and motivate proper scoring rules with examples from Bayesian probability and robust statistics. We show that the need to optimize unnormalized probability densities, λp_θ, in some estimation

problems with respect to both $\lambda > 0$ and $\theta \in \Theta$ naturally leads to the notion of a proper affine scoring rule.

Section 3 is dedicated to describing the characteristic properties of Bregman divergences. We first consider Bregman divergences in the best-known context of Euclidean spaces. We begin with an overview of the basic mathematical properties of Bregman divergences such as symmetry and joint and separate convexity [5]. We then discuss the fact that Bregman divergences may also be characterized probabilistically: They are the unique class of divergences with the property that the expected divergence of every random variable to a given value is minimized by the mean of the random variable [2,32]. The result remains true for multivariate variables and is invariant with respect to any transformation of these variables [1,14]. A divergence characterizes as Bregman if and only if it is a convex function in the true distribution inducing the same Bregman divergence [3]. We generalize this result to the current measure-theoretic framework. Some generalizations of Bregman divergence to function spaces have been investigated by [15], however, the authors do not work in the general framework of spaces associated with scoring rules. We adapt to our framework the result of [17] which characterizes the proper affine scoring rules as the affine component of a functional Bregman divergence on a general (nonprobabilistic) domain. Combining the results of this section, we resolve the problem stated in [19], Section 2.2, to classify all score divergences.

In Section 4, we consider proper scoring rules only on domains of normalized probability distributions. The results presented here are mostly classical and serve as a basis for the more recent generalizations in Section 3.2. We first give the direct characterization of proper scoring rules on the probability simplex due to [19,32]. We then present another characterization of properness due to [27], who shows that the latter is equivalent to a seemingly stronger condition that we term order sensitivity, borrowing its name from [24,33]. Specifically, a scoring rule is order sensitive if its expected score increases whenever the predictive distribution moves in the direction of the true distribution. Lastly, we present the earliest characterization of proper scoring rules due to [20,26]. The result characterizes proper scoring rules on the positive orthant as subgradients of 1-homogeneous convex functions.

In Section 5, our goal is to understand under what conditions a general convex function constitutes the entropy of a proper scoring rule and whether the association between entropy and proper scoring rule is unique. This requires knowledge of the regularity properties of convex functions. The latter may be described simply in finite dimensions, where every convex function on an open domain is the entropy function of a unique proper scoring rule up to a potentially negligible set of distributions. In infinite dimensions, however, the regularity theory of entropy functions comes in two flavours depending on whether the entropy is continuous or discontinuous. The first case is well-known and similar to that in the Euclidean setting, while the second case requires some less-known concepts such as the quasi-interior of a convex set. The discontinuous case is exemplified by the Shannon and Hyvärinen entropies. We studied it in [29] under the assumption that the entropy function is 1-homogeneous and presented necessary and sufficient conditions for existence and uniqueness of subgradients that may be identified as proper scoring rules. We also gave examples of domains on which the logarithmic and Hyvärinen scoring rules are the unique subgradients of their 1-homogeneous entropies. Here, we extend these results to their general form by dropping the assumption that the entropy is 1-homogeneous.

2. Preliminaries

We begin by introducing the notation and underlying mathematical framework. We then give the formal definitions of a proper scoring rule, an entropy function, and a divergence function. In Section 2.3, we illustrate the role which proper scoring rules have played in justifying the notions of probabilistic coherence and Bayesian probability. In Section 2.4, we consider an example from robust estimation motivating the need to formalize the proper affine scoring rules.

2.1. Notation

We consider a probability space $(\Omega, \mathcal{A}, \mathsf{P})$, a measurable space $(\mathcal{X}, \mathcal{B})$, and a random variable $X : (\Omega, \mathcal{A}, \mathsf{P}) \to (\mathcal{X}, \mathcal{B})$. We assume that $\mathcal{X}$ is a subset of a Euclidean space and $\mathcal{B}$ is the associated Borel σ-algebra. Here X plays the role of an observed quantity for which a forecaster makes probabilistic predictions. The latter are probability distributions on $(\mathcal{X}, \mathcal{B})$ that aim to represent the uncertainty in the range of possible outcomes of X. Probabilistic forecasts are more general and more informative than single-valued forecasts and may always be converted into the latter by taking the mean, the mode, etc., of the predictive distribution. By $\mathcal{P}$ we denote a convex set of probability measures on $(\mathcal{X}, \mathcal{B})$. The set contains all feasible distributions for X, including its true distribution. We use the term probability measure and the associated cumulative distribution function interchangeably. Let $\operatorname{cone} \mathcal{P} = \{\lambda P | \lambda > 0, P \in \mathcal{P}\}$ denote the convex cone of $\mathcal{P}$. By $\operatorname{span} \mathcal{P}$, we denote the linear span of $\mathcal{P}$, that is, the collection of all finite linear combinations of elements in $\mathcal{P}$.

Definition 2.1. We call the functions $f : \mathcal{X} \to \mathbb{R}$ $\mathcal{P}$*-integrable* if f is measurable[1] and

$$\int_{\mathcal{X}} \left| f(x) \right| dP(x) < \infty \tag{1}$$

for every $P \in \mathcal{P}$. We denote by $\mathcal{L}(\mathcal{P})$ the linear space of all $\mathcal{P}$-integrable functions.

Let us notice that the spaces $\operatorname{span} \mathcal{P}$ and $\mathcal{L}(\mathcal{P})$ are dual and every $f \in \mathcal{L}(\mathcal{P})$ can be viewed both as a function on $\mathcal{X}$ and as a linear functional on $\operatorname{span} \mathcal{P}$. By "$\cdot$" we denote the duality operation between $\operatorname{span} \mathcal{P}$ and $\mathcal{L}(\mathcal{P})$. In more abstractly orientated treatments of the subject scoring rules are defined as function-valued maps of the form $\mathbf{S} : \mathcal{P} \to \mathcal{L}(\mathcal{P})$. For every $P \in \mathcal{P}$, $\mathbf{S}(P)$ can be viewed both as a linear functional acting on the space of finite signed measures $\operatorname{span} \mathcal{P}$ and as a random function $\mathbf{S}(P)(X)$ of X.[2] In most of the modern literature, however, scoring rules are more simply defined as functions of the form $S : \mathcal{P} \times \mathcal{X} \to \mathbb{R}$ such that, for every $P \in \mathcal{P}$, $S(P, \cdot)$ is a $\mathcal{P}$-integrable function. The quantity $S(P, x)$ denotes the score assigned to a forecast $P \in \mathcal{P}$

[1] In this context it is not essential to distinguish between Borel and Lebesgue measurability of f. This follows from the fact that if f is Borel-measurable then it is also Lebesgue-measurable, while if f is Lebesgue-measurable then there is a Borel-measurable function g such that $f = g$ almost everywhere.

[2] This interpretation of scoring rules and the "$\cdot$" notation was introduced by [20].

and a subsequent observation $x \in \mathcal{X}$. The two definitions agree by setting $S(P, x) = \mathbf{S}(P)(x)$. We write

$$S(P, Q) = \mathbb{E}_{X\sim Q} S(P, X) = \int_{\mathcal{X}} S(P, x)\, dQ(x) = \mathbf{S}(P) \cdot Q$$

for the *expected score* of the forecast P under the true distribution Q of X.[3]

In a number of situations we need to consider scoring rules outside the set of probabilities $\mathcal{P}$. In game-theoretic setting, this has been theorized by [17], while here we develop the idea purely within the realm of statistics. We thus extend the domain of a scoring rule to the positive orthant, $\operatorname{cone}\mathcal{P}$, which necessitates an extension of the range of the scoring rule to the space of all affine $\mathcal{P}$-integrable functionals on $\operatorname{span}\mathcal{P}$. The latter is denoted by $\mathcal{A}(\mathcal{P})$ and we have that $l \in \mathcal{A}(\mathcal{P})$ if and only if l is in the form $l(P) = a^* \cdot P + \alpha$, where $P \in \operatorname{span}\mathcal{P}$ is arbitrary, while $a^* \in \mathcal{L}(\mathcal{P})$ and $\alpha \in \mathbb{R}$ are fixed. Any mapping $\mathbf{S} : \operatorname{cone}\mathcal{P} \to \mathcal{A}(\mathcal{P})$ is said to be an *affine scoring rule* on $\operatorname{cone}\mathcal{P}$. Unlike ordinary scoring rules, affine scoring rules are not naturally defined for single observations. Instead, affine scoring rules are applied only to probability distributions; we write $\mathbf{S}(P)(Q)$ for the *expected score* of P under Q, where $P \in \operatorname{cone}\mathcal{P}$ and $Q \in \operatorname{span}\mathcal{P}$ are arbitrary.[4,5]

In almost all cases of practical importance, the distributions P, Q are represented by probability densities with respect to a fixed measure μ on $(\mathcal{X}, \mathcal{B})$. As a rule, the measure μ is either the counting measure when $\mathcal{X}$ is discrete, or the Lebesgue measure when $\mathcal{X}$ is an open set (in a Euclidean space). We may unite the two cases by referring to a probability mass function as a probability density with respect to the counting measure. Probability densities are always denoted by lower-case letters p, q, while their cumulative distribution functions are denoted by upper-case letters P, Q, respectively. Whenever the distributions in $\mathcal{P}$ have densities, we associate them with their densities. In this case, we write

$$S(p, q) := \mathbb{E}_{X\sim q} S(p, X) = \int_{\mathcal{X}} S(p, x) q(x)\, d\mu(x) = \mathbf{S}(p) \cdot q, \qquad p, q \in \mathcal{P},$$

for the expected score of p under q.

Example 2.2. If the sample space $\mathcal{X} = \{x_1, \ldots, x_d\}$ is discrete and has d elements, then we take by default $\mathcal{P}$ to be the probability simplex in $\mathbb{R}^d$, while $\operatorname{cone}\mathcal{P}$ is the positive orthant in $\mathbb{R}^d$. In addition, $\operatorname{span}\mathcal{P}$ is the Euclidean space $\mathbb{R}^d$ and its dual space $\mathcal{L}(\mathcal{P})$ is identical to $\mathbb{R}^d$, too.

Let a function $\Phi : \operatorname{cone}\mathcal{P} \to \mathbb{R}$ be given. It is said that Φ is (positively) α-*homogeneous* for some $\alpha \in \mathbb{R}$ if $\Phi(\lambda a) = \lambda^\alpha \Phi(a)$ for every $a \in \mathcal{P}$ and every $\lambda > 0$. Let $\mathcal{X} \subset \mathbb{R}^d$ be an open set and $1 \le p < \infty$. By $L^p(\mathcal{X})$, we denote the Lebesgue L^p-space of functions on $\mathcal{X}$ whose pth

[3] For the letters denoting the true and predictive distribution and their order of appearance in the expected score we follow the convention established in the probabilistic forecasting community [11,19]. This convention is likely based on the fact that in diagrams visualizing the calibration of probabilistic forecasts it is natural to use the horizontal axis to display the forecasts and the vertical axis to display the observations.

[4] Notice that on $\mathcal{P}$ every affine scoring rule reduces to an ordinary scoring rule.

[5] Non-affine generalizations of proper scoring rules were considered by [22]. In particular, they introduce a continuum of generalized scoring rules that connects the power and pseudospherical scoring rules.

power of the absolute value is Lebesgue integrable. The associated L^p-norm is denoted by $\|\cdot\|_p$. We use upper indices to denote sequences of observations in $\mathcal{X}$, as in x^i. An element $x \in \mathcal{X}$ has component form $x = (x_1, \ldots, x_d)$ and $|x| = (|x_1| + \cdots + |x_d|^2)^{\frac{1}{2}}$ is its magnitude. We sometimes write the partial derivative $\partial/\partial x_i$, $i = 1, \ldots, d$, as ∂_i. The gradient on $\mathcal{X}$ is denoted by

$$\nabla = \left(\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_d} \right).$$

2.2. Definition of a proper scoring rule

Proper scoring rules are protocols for eliciting and evaluating probabilistic forecasts. By being maximized in expectation at the true prediction, they incentivize a forecaster to truthfully report his private information.

Definition 2.3. A scoring rule $S : \mathcal{P} \times \mathcal{X} \to \mathbb{R}$ that maximizes its expected score,

$$S(Q, Q) = \max_{P \in \mathcal{P}} S(P, Q), \tag{2}$$

at the true distribution $Q \in \mathcal{P}$ is called *proper*. If the true distribution is always a unique maximizer, S is called *strictly proper*.

The definition of a proper scoring rule is equivalent to the condition that $S(P, Q) \leq S(Q, Q)$, for all $P, Q \in \mathcal{P}$, and S is strictly proper, if the latter inequality is strict for all $P \neq Q$. The function $\Phi : \mathcal{P} \to \mathbb{R}$ given by $\Phi(P) = S(P, P)$, for every $P \in \mathcal{P}$, is the *(negative) entropy function* associated with S. Notice that in view of (2), Φ must be convex on $\mathcal{P}$ as a pointwise maximum of linear functionals. Moreover, Φ is strictly convex if and only if S strictly proper. Consider the function

$$D_S(P, Q) = S(Q, Q) - S(P, Q), \qquad P, Q \in \mathcal{P}. \tag{3}$$

The condition $D_S(P, Q) \geq 0$ for all $P, Q \in \mathcal{P}$ is equivalent to S being proper, while the additional requirement that $D_S(P, Q) = 0$ if and only if $P = Q$ is equivalent to S being strictly proper.

Definition 2.4. A function $D : \mathcal{P} \times \mathcal{P} \to [0, \infty)$ is called a *divergence* on $\mathcal{P}$. If additionally $D(P, Q) = 0$ if and only if $P = Q$, then D is called a *strict divergence*.

We conclude that S is a (strictly) proper scoring rule if and only if the associated function D_S is a (strict) divergence. A divergence D_S induced by a scoring rule is called a *score divergence*. Many authors prefer to define divergences as what we call strict divergences because only the latter guarantee consistency in estimation problems of the associated estimators. In the context of score divergences, however, we prefer the current more nuanced definition. In the same fashion, we may also differentiate between strict and non-strict entropy functions. We shall later show in

Section 3 that every score divergence may be identified as a functional Bregman divergence on a subset of the space of finite signed measures, span $\mathcal{P}$.

In the context of an affine scoring rule $\mathbf{S} : \operatorname{cone} \mathcal{P} \to \mathcal{A}(\mathcal{P})$, Definition 2.3 is modified accordingly: $\mathbf{S}$ is proper if, for every $P, Q \in \operatorname{cone} \mathcal{P}$, the expected score $\mathbf{S}(P)(Q)$ is maximized in P at the true distribution Q. If the expected score is uniquely maximized at the true distribution, then $\mathbf{S}$ is called strictly proper. The definition of a divergence on domains larger than $\mathcal{P}$ is analogous to Definition 2.4. Proper affine scoring rules define score divergences analogously to proper scoring rules.

2.3. Formalizing Bayesian probability

This subsection contains a recent take on a classical subject and has mostly illustrative purposes.

To begin with, we consider a binary predicted variable X taking values in the sample space $\mathcal{X} = \{0, 1\}$. The variable X may be seen as the indicator function

$$1_E(\omega) = \begin{cases} 1, & \omega \in E, \\ 0, & \omega \notin E \end{cases}$$

of the event $E = \{\omega \in \Omega | X(\omega) = 1\}$. The distribution of X is a vector with components $\mathsf{P}(X = 0) = 1 - p$ and $\mathsf{P}(X = 1) = p$, with $p \in [0, 1]$. The endpoint cases $p = 0$ and $p = 1$ correspond to a constant X and often may be excluded without loss of generality. We may represent the probability distribution of X by a single number, which by convention we take to be $\mathsf{P}(X = 1) = p$. This allows us to interpret a probabilistic forecast for X as a probability forecast for the event E.

Examples of binary proper scoring rules are given by the *log score*,

$$S(p, i) = \ln p_i,$$

and the *Brier score*,

$$S(p, i) = -(i - p_i)^2,$$

where $p = (p_0, p_1)$ is a probability distribution and $p_i = \mathsf{P}(X = i)$.

We now proceed to discuss what it means for a person to have logically consistent beliefs about a collection of events. Let $E_i \in \mathcal{A}$, $i = 1, \ldots, n$, be n events in Ω, which we denote as the vector $E = (E_1, \ldots, E_n)$. By a vector forecast for E we understand any vector $p = (p_1, \ldots, p_n)$, $p_i \in [0, 1]$, such that p_i is the probability forecast prescribed to the event E_i. Let us fix a strictly proper scoring rule S and a vector of events E. The *combined score* of a vector forecast p under the outcome $\omega \in \Omega$ is

$$S_{\mathrm{com}}(p, \omega) = \sum_{i=1}^{n} S\big(p_i, 1_{E_i}(\omega)\big).$$

This is just a sum of scores assigned to n probability forecasts under the same realization ω. Here, the distribution for each variable $X_i = 1_{E_i}$ is represented by a single number, p_i, rather than the full vector $(1 - p_i, p_i)$.

Definition 2.5. A vector forecast p is *dominated* by a vector forecast q if $S_{\text{com}}(p,\omega) \leq S_{\text{com}}(q,\omega)$, for all $\omega \in \Omega$, and *strongly dominated* by q if the inequality is strict for all $\omega \in \Omega$. A vector forecast p is called *admissible* if it is not dominated by a vector forecast other than itself.

In the framework of Bayesian probability, only the admissible vector forecasts are deemed as rational representations of beliefs. This is based on the idea that beliefs may be expressed through the betting behaviour of an individual in lottery games and an inadmissible vector forecast leads to a sure-loss bet. See, for example, [13]. We are next going to characterize the class of admissible vector forecasts.

Definition 2.6. We call a vector forecast $p = (p_1, \ldots, p_n)$ for the events $E = (E_1, \ldots, E_n)$ *probabilistically coherent* whenever there is a probability measure ν over $(\Omega, \mathcal{A})$ such that $\nu(E_i) = p_i$.

Under mild regularity conditions on the scoring rule, [30] show the following.

Theorem 2.7. *The admissible vector forecasts are precisely those that are probabilistically coherent. Moreover, any incoherent vector forecast is strongly dominated by some coherent vector forecast.*

To illustrate the theorem, consider , for example, events $E = (E_1, E_2)$ such that $E_1 \subseteq E_2$. Then, for any vector forecast $p = (p_1, p_2)$ for E that does not satisfy the inequality $p_1 \leq p_2$, there is a another vector forecast that strongly dominates it. The result lends support to the idea that any rational description of uncertainty by numbers must obey the rules of probability calculus. The proof of Theorem 2.7 relies on the notion of Bregman divergence and provides a new insight into classical results obtained by [25,32].

2.4. On the estimation of unnormalized probability densities

Here, we consider an application of scoring rules to robust statistics which will motivate their extension as affine scoring rules on cone $\mathcal{P}$. Our example involves the *power scoring rule* $S_{\text{pow}} : \mathcal{P} \times \mathcal{X} \to \mathbb{R}$, given by

$$S_{\text{pow}}(p,x) = \gamma p(x)^{\gamma-1} - (\gamma-1)\|p\|_\gamma^\gamma,$$

and the *pseudospherical scoring rule* $S_{\text{psu}} : \mathcal{P} \times \mathcal{X} \to \mathbb{R}$, given by

$$S_{\text{psu}}(p,x) = \lambda \frac{p(x)^{\gamma-1}}{\|p\|_\gamma^{\gamma-1}},$$

both defined for $\gamma > 1$.

Suppose we draw a sample from a contaminated probability density

$$q_\varepsilon(x) = (1-\varepsilon)q(x) + \varepsilon r(x), \tag{4}$$

where q is the true or target density, r is the contamination density, and $\varepsilon \in [0,1]$ is the rate of contamination. A standard estimation method tries placing the model density close to q_ε, which is not favourable as our real goal is estimating q, not q_ε. A robust estimation method tries instead placing the model density close to q. We consider (4) in the case where $q \sim N(0,1)$ and $r \sim N(5,1)$ are normal densities and 20% of the sample is contaminated, corresponding to $\varepsilon = 0.2$. Let us compare how well the power and pseudospherical scoring rules are able to estimate the target density in this context.

To that end, let us suppose that the mean $\mu = 0$ of q is known and try to estimate its standard deviation σ. Thus, we employ the parametric model $p_\sigma \sim N(0,\sigma^2)$, where

$$p_\sigma(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}$$

is the univariate Gaussian. We draw from (4) an independent and identically distributed sample $x_1, \ldots, x_n$, where $n = 1000$. The estimating equation of the power scoring rule is

$$c(\gamma)\sigma + \frac{1}{n}\sum_{i=1}^{n}\big(-\sigma^2 + (x_i-\mu)^2\big)p_{\sigma/\sqrt{\gamma-1}}(x_i) = 0, \tag{5}$$

where

$$c(\gamma) = \left(\frac{\gamma-1}{\gamma}\right)^{\frac{3}{2}} \frac{1}{\sqrt{2\pi}}, \qquad \gamma > 1. \tag{6}$$

The estimating equation of the pseudospherical scoring rule is

$$\sum_{i=1}^{n}\big(-\sigma^2/\gamma + (x_i-\mu)^2\big)p_{\sigma/\sqrt{\gamma-1}}(x_i) = 0. \tag{7}$$

The above two estimating equations have been adapted to our context from the more general equations in [4,18], or in [12], equation (13). We solve (5) and (7) in σ to find the corresponding score estimators. We repeat this 100 times and find the average values of each estimator. From the two average values, we subtract the true value of σ ($\sigma = 1$) and obtain an estimate of the bias of each estimator. The results are displayed in Figure 1. The graph exhibits the bias of each estimator in the range $\gamma \in [1,2]$. We see that the power scoring rule has a persistent bias across all γ, while the pseudospherical scoring rule, on the other hand, is approximately unbiased for a sufficiently large γ.

We now try to explain the apparent difference in robustness between the two scoring rules. To that end, let Θ be the parameter space of a given model and assume that the contamination density r lies in the tails of the model density p_θ so that the quantity

$$\varepsilon_\theta = p_\theta^{\gamma-1} \cdot r$$

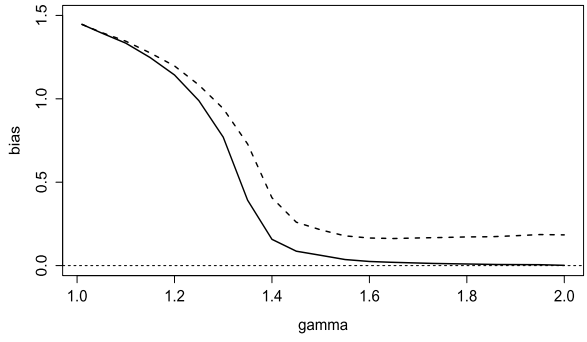

Figure 1. Bias of the pseudospherical (solid lines) and power scoring rule (dashed lines) for $\gamma \in [1, 2]$ in the estimation of the standard deviation σ of the true density q in (4).

is sufficiently small for an appropriately large $\gamma > 1$ and all $\theta \in \Theta$. Notice that the function $\lambda \to S_{\mathrm{psu}}(p_\theta, \lambda q)$ is linear in $\lambda > 0$, where

$$S_{\mathrm{psu}}(p_\theta, \lambda q) = \lambda \frac{p_\theta^{\gamma-1} \cdot q}{\|p_\theta\|_\gamma^{\gamma-1}} = \lambda S_{\mathrm{psu}}(p_\theta, q)$$

is the expected score of the pseudospherical scoring rule. As a result of this, we have that

$$S_{\mathrm{psu}}\big(p_\theta, (1-\varepsilon)q + \varepsilon r\big) = (1-\varepsilon) S_{\mathrm{psu}}(p_\theta, q) + \varepsilon c_\theta \varepsilon_\theta,$$

where $c_\theta = \|p_\theta\|_\gamma^{1-\gamma}$. Assuming that the term $\varepsilon c_\theta \varepsilon_\theta$ is sufficiently small, we may neglect it and obtain that

$$\arg\max_{\theta\in\Theta} S_{\mathrm{psu}}(p_\theta, q_\varepsilon) \approx \arg\max_{\theta\in\Theta} S_{\mathrm{psu}}(p_\theta, q),$$

which explains the robustness of the pseudospherical scoring rule under heavy contamination. On the other hand, the function $\lambda \to S_{\mathrm{pow}}(p_\theta, \lambda q)$ is affine, where

$$S_{\mathrm{pow}}(p_\theta, \lambda q) = \lambda\gamma p_\theta^{\gamma-1} \cdot q - (\gamma-1)\|p_\theta\|_\gamma^\gamma,$$

is the expected score of the power scoring rule. As above, we have that

$$S_{\mathrm{pow}}\big(p_\theta, (1-\varepsilon)q + \varepsilon r\big) = S_{\mathrm{pow}}\big(p_\theta, (1-\varepsilon)q\big) + \gamma\varepsilon\varepsilon_\theta,$$

which shows the approximate equivalence of the optimization problems

$$\arg\max_{\theta\in\Theta} S_{\mathrm{pow}}(p_\theta, q_\varepsilon) \approx \arg\max_{\theta\in\Theta} S_{\mathrm{pow}}\big(p_\theta, (1-\varepsilon)q\big).$$

However, the latter problem is clearly not approximately equivalent to the desired one, $\arg\max_{\theta\in\Theta} S_{\mathrm{pow}}(p_\theta, q)$, unless $\varepsilon > 0$ is sufficiently small, due to the nonlinearity (affinity) of the expected score of S_{pow}. This heuristically explains the non-robustness of the power scoring rule under heavy contamination. The above analysis and example have been adapted from [18].

Notwithstanding the preceding analysis, it turns out that the deficiency in robustness of S_{pow} in comparison to S_{psu} may be completely overcome if the optimization problem is formulated in

a suitable way. Indeed, let us notice that

$$S_{\text{pow}}\big(\lambda p_\theta, (1-\varepsilon)q\big) = (1-\varepsilon)^\gamma S_{\text{pow}}(p_\theta, q)$$

for $\lambda = (1-\varepsilon)$. This motivates the model

$$\{\lambda p_\theta | \lambda > 0, \theta \in \Theta\},$$

called an *enlarged parametric model*, which we use to formulate the two-fold optimization problem

$$(\hat{\lambda}, \hat{q}) = \max_{\lambda \in [0,1], \theta \in \Theta} S_{\text{pow}}(\lambda p_\theta, q_\varepsilon). \tag{8}$$

Under mild regularity assumptions, [23] show that $\hat{\lambda}$ is an approximately unbiased estimator for the target ratio, $1-\varepsilon$, and $\hat{q}$ is an approximately unbiased estimator for the target density, q. Moreover, the latter estimator turns out to be identical to the score estimator of S_{psu}, thus showing a surprising equivalence between the two scoring rules [23], Theorem 1.

The above example and analysis indicate the need to consider enlarged parametric models, where the model, empirical, and population distributions may not be normalized. This motivates the formal introduction of proper affine scoring rules.

3. Bregman divergence

Although the notions of Bregman divergence and proper scoring rules are intimately related, their origin and development occurred to a large extent separately. Indeed, the former has been used primarily in convex optimization as a generalization of the Euclidean distance, while the latter are protocols for eliciting and evaluating probabilistic forecasts. By developing Bregman divergences in the framework of spaces of finite signed measures, we may identify score and Bregman divergences. We begin with an overview of the characteristic properties of Bregman divergence in Euclidean spaces, and then address the analogous question in the context of arbitrary spaces of finite signed measures.

3.1. Bregman divergence in Euclidean spaces

Let I be an open interval in $\mathbb{R}$ and consider a convex function $\phi : I \to \mathbb{R}$ that is continuously differentiable. Let also $\phi'(a)$ denote the derivative of ϕ at a.

Definition 3.1. The function $d_\phi : I \times I \to [0, \infty)$, given by

$$d_\phi(a, x) = \phi(x) - \phi'(a)(x-a) - \phi(a) \tag{9}$$

is the *Bregman divergence* associated with ϕ.[6]

[6]The order and meaning of the arguments is reversed with respect to the convention used for example, in information theory, but is in line with that of probabilistic forecasting [11,19].

Geometrically, $d_\phi(a,x)$ is nonnegative whenever the graph of ϕ lies above its tangent line at $x=a$. Such a tangent line is said to *support* ϕ at a. Moreover, a continuously differentiable function is convex if and only if the function lies above all of its tangents. If in addition we have equality in $d_\phi(a,x)\geq 0$ precisely for $x=a$, then the supporting lines of ϕ have a single point of tangency with the graph of ϕ. The latter is equivalent to ϕ being strictly convex.

Example 3.2. One-dimensional Bregman divergences are associated with an important class of divergences for probability densities. Let again $\mathcal{P}$ be a convex set of probability densities with respect to a fixed measure μ on the sample space $\mathcal{X}$. If the domain of ϕ is $I=(0,\infty)$, we may construct the divergence $D_\phi:\mathcal{P}\times\mathcal{P}\to[0,\infty)$ by setting

$$D_\phi(p,q)=\int_{\mathcal{X}} d_\phi\big(p(x),q(x)\big)\,d\mu(x)$$

for any $p,q\in\mathcal{P}$. Divergences of that form are called *separable* because they are sums or integrals of a fixed scalar divergence applied pointwise to p and q. Associated with every separable Bregman divergence d_ϕ is a proper scoring rule, $S_\phi:\mathcal{P}\times\mathcal{X}\to\mathbb{R}$, defined as

$$S_\phi(p,x)=\phi'\big(p(x)\big)-\int_{\mathcal{X}}\big(\phi'\big(p(x)\big)p(x)-\phi\big(p(x)\big)\big)\,d\mu(x),$$

a convex entropy function $\Phi:\mathcal{P}\to\mathbb{R}$, given by

$$\Phi(p)=\int_{\mathcal{X}}\phi\big(p(x)\big)\,d\mu(x),$$

and we have that

$$D_\phi(p,q)=\Phi(q)-S_\phi(p,q),\qquad \text{for every } p,q\in\mathcal{P}.$$

For example, the logarithmic and power scoring rule are associated with separable Bregman divergences, while the pseudospherical scoring rule is not.

We now describe three basic mathematical properties of Bregman divergences of the form (9). First, $d_\phi(a,x)$ is called *symmetric* whenever $d_\phi(a,x)=d_\phi(x,a)$ for all $a,x\in I$. Symmetry is a special property and holds only for the quadratic divergence[7] $d(a,x)=(x-a)^2$ [5,32]. Second, it is clear that $d_\phi(a,x)$ is convex in x for every $a\in I$, so in addition it is reasonable to consider the following two stronger properties:

(i) d_ϕ is *jointly convex* if $(a,x)\to d_\phi(a,x)$ is convex on $I\times I$;
(ii) d_ϕ is *separately convex* if $a\to d_\phi(a,x)$ is convex on I, for every $x\in I$.

Joint convexity implies separate convexity but not conversely [5]. The main result of the latter work is to show that d_ϕ is jointly convex if and only if $1/\phi''$ is concave, provided that $\phi''(x)$ exists

[7] We identify all divergences differing by a multiplicative constant.

and is positive for every $x \in I$. One may verify that for example $\phi(x) = x \ln x$ and $\phi(x) = x^\gamma$, for $1 < \gamma \leq 2$ satisfy this criterion on $I = (0, \infty)$. Separate convexity is characterized in [5], Theorem 3.11.

We now proceed to describe Bregman divergences in $\mathbb{R}^d$.

Definition 3.3. Let $\phi : U \to \mathbb{R}$ be a continuously differentiable convex function defined on an open convex domain $U \subset \mathbb{R}^d$. The function $d_\phi : U \times U \to [0, \infty)$,

$$d_\phi(a, x) = \phi(x) - \nabla\phi(a) \cdot (x - a) - \phi(a), \tag{10}$$

is the *Bregman divergence* associated with ϕ.

In fact, (10) is the remainder evaluated at x of the first-order Taylor series expansion of ϕ around a. The following result is a mathematical folklore [5].

Lemma 3.4. *Let d_ϕ be defined as in* (10)*, where $\phi : U \to \mathbb{R}$ is an arbitrary continuously differentiable function. Then,*

(i) $d_\phi(a, x) \geq 0$ *(with equality only for $x = a$) if and only if ϕ is (strictly) convex*;
(ii) $d_\phi(a, x) = 0$ *for all $x, a \in U$ if and only if ϕ is affine.*

Thus, two convex functions define the same Bregman divergence if and only if their difference is an affine function. The properties of symmetry, joint convexity, and separate convexity are defined analogously to the one-dimensional case. Let A be a positive semi-definite matrix of dimension d. The bilinear form

$$\phi(x) = x^t A x$$

is a convex function. Furthermore, [6] show that the associated *generalised quadratic divergence*,

$$d_A(p, q) = (p - q)^t A (p - q),$$

closely related to Mahalanobis distance, is the only symmetric Bregman divergence on $\mathbb{R}^d$. Evidently, d_A is separable if and only if the matrix A is diagonal. For characterization of joint convexity see [5], Theorem 6.1.

Example 3.5. Perhaps a bit surprisingly, Bregman divergences may also be characterized probabilistically. Indeed, they are the unique class of divergences for which the mean divergence of a random variable to a fixed value is minimized by the mean of the random variable. Specifically, let $U \subset \mathbb{R}^d$ be open and convex and $d_0 : U \times U \to [0, \infty)$ be such that $d_0(a, x) = 0$ if and only if $a = x$, for all $a, x \in U$. If for all random variables X taking values in U we have that

$$\mathbb{E}X = \arg\min_{a \in U} \mathbb{E}\, d_0(a, X),$$

and d_0 satisfies mild regularity conditions, then there is a continuously differentiable strictly convex function $\phi : U \to \mathbb{R}$ such that $d_0 = d_\phi$. Conversely, if d_0 has the form d_ϕ then the above

minimization property holds.[8] The result is best known in the special case where $d_0(x, y) = (x - y)^2$ is the squared error. See, for example, [2,32]. The assertion is invariant with respect to transformations of X. Let, for example, $g : U \to V$ be a continuously differentiable and invertible transformation, where V is also open and convex set in $\mathbb{R}^d$. Then, if $s : U \times V \to [0, \infty)$ is such that for all random variables X taking values in U we have that

$$\mathbb{E}g(X) = \arg\min_{a \in V} \mathbb{E}s(a, X),$$

and s is subject to mild regularity assumptions, then there is a continuously differentiable strictly convex function $\psi : V \to \mathbb{R}$ such that $s(a, x) = d_\psi(a, g(x))$. For proof of the claim in both directions see [16], Section 4.2.1, or [1,14].

Bregman divergences may also be defined for arbitrary convex functions by replacing the gradient with a selection of subgradients. This is discussed more generally in the next subsection. An overview of the basic regularity properties of convex functions in Euclidean spaces may be found in Section 5.1.

3.2. Functional Bregman divergence

Here we present Bregman divergences in the context of spaces of finite signed measures. To that end, let $(\mathcal{X}, \mathcal{B})$ be a measurable space such that $\mathcal{X}$ is a subset of an Euclidean space and $\mathcal{B}$ is the associated Borel σ-algebra. The set $\mathcal{P}$ is a convex subset of the set of all probability measures on $(\mathcal{X}, \mathcal{B})$. The vector space $\operatorname{span}\mathcal{P}$ is the linear span of $\mathcal{P}$, while $\mathcal{L}(\mathcal{P})$ is the dual space to $\operatorname{span}\mathcal{P}$ and is defined as the collection of all $\mathcal{P}$-integrable functions on $\mathcal{X}$. In what follows, we fix a convex set U such that $\mathcal{P} \subset U \subset \operatorname{span}\mathcal{P}$.

In order to keep the presentation sufficiently general, we are not going to assume that the set U is open and that the convex functions on U are continuous or differentiable. To motivate this, we consider the following.

Example 3.6. Let $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{N}$ be a linear subspace of $L^1(\mathcal{X})$ containing continuous functions. By $\mathcal{P}$ we denote the set of all probability densities in $\mathcal{N}$ that are strictly positive. We are going to show that the positive orthant in $\mathcal{N}$, denoted as $\operatorname{cone}\mathcal{P}$, has empty algebraic interior in $\mathcal{N}$ if the elements of $\operatorname{cone}\mathcal{P}$ satisfy the following condition

$$\forall p \in \operatorname{cone}\mathcal{P}\ \exists x_0 \in \mathcal{X}\ \exists q \in \operatorname{cone}\mathcal{P} : \lim_{x \to x_0} \frac{p(x)}{q(x)} = 0,\ ^{9} \tag{11}$$

where x_0 could also denote infinity if $\mathcal{X}$ is unbounded. For example, this condition will be satisfied if we take $\mathcal{X} = \mathbb{R}$, $x_0 = \infty$, and two Gaussians p and q on the real line such that p has

[8] Analogous claim holds for non-strict Bregman divergences.

[9] Notice that this condition cannot hold in finite dimensions and signifies a qualitative difference between the finite and infinite-dimensional settings.

smaller variance than q. Let condition (11) hold and notice that the function $p(x) - tq(x)$ cannot remain positive as $x \to x_0$, no matter how small $t > 0$ is. It is clear that p cannot be a core point for cone $\mathcal{P}$ and thus the latter has empty algebraic interior in $\mathcal{N}$. (See [35], page 2, or [28], Section A.3, for definitions.) The implications of this fact are that any entropy function, such as the Shannon or Hyvärinen entropy, that cannot be extended past the positive orthant cone $\mathcal{P}$ will not be continuous or differentiable, regardless of the choice of topology on $\mathcal{N}$.

The example shows that in the context of continuous variables there are important entropy functions that are not continuous, differentiable, or even finite on the open domains of the normed subspaces of $L^1(\mathcal{X})$. Local regularity properties of entropy functions like these will be the subject of Section 5. The existence of subgradient, however, is a global property that is well-defined in the context of a general vector space. The exposition here will be based on the latter property and will be independent of topology.

Definition 3.7. If $\Phi : U \to \mathbb{R}$ and there is $P^* \in \mathcal{L}(\mathcal{P})$ such that

$$\Phi(Q) \geq P^* \cdot (Q - P) + \Phi(P), \qquad \text{for all } Q \in U, \tag{12}$$

then we say that P^* is a subgradient of Φ at P. If the inequality is strict for all $P \neq Q$, then P^* is called a strict subgradient.

Geometrically, the inequality in (12) means that the graph of Φ is supported at P by the hyperplane

$$\pi_P = \big\{(Q, y) | y = P^* \cdot (Q - P) + \Phi(P)\big\} \tag{13}$$

in the vector space $(\operatorname{span} \mathcal{P}, \mathbb{R})$. The collection of all subgradients of Φ at P is called the *subdifferential* of Φ at P and denoted by $\partial\Phi(P)$. Suppose that $\partial\Phi(P) \neq \varnothing$ for each $P \in U$. Then, we call a selection of subgradients $\Phi^*(P) \in \partial\Phi(P)$, for each $P \in U$, a *subgradient* of Φ on U. In view of the fact that $\Phi^*(P) \in \mathcal{L}(\mathcal{P})$, we may identify $\Phi^*(P)$ with its $\mathcal{P}$-integrable kernel function on $\mathcal{X}$, denoted as $\Phi^*(P)(x)$.

Definition 3.8. Suppose that $\Phi : U \to \mathbb{R}$ has a subgradient $\Phi^* : U \to \mathcal{L}(\mathcal{P})$. Then the function $D_{(\Phi,\Phi^*)} : U \times U \to \mathbb{R}$ given by

$$D_{(\Phi,\Phi^*)}(P, Q) = \Phi(Q) - \Phi^*(P) \cdot (Q - P) - \Phi(P), \tag{14}$$

for all $P, Q \in U$, is the *functional Bregman divergence* on U associated with the pair (Φ, Φ^*).

Geometrically, $D_{(\Phi,\Phi^*)}(P, Q)$ is the vertical distance at Q between the graph of Φ and its supporting hyperplane at P induced by $\Phi^*(P)$. As a mnemonic rule, the forecast distribution P selects the supporting hyperplane of Φ, while the true distribution Q determines the place of vertical distance. The next result is proved analogously to Lemma 3.4.

Lemma 3.9. *Suppose that $\Phi : U \to \mathbb{R}$ has a subgradient $\Phi^* : U \to \mathcal{L}(\mathcal{P})$. If $D_{(\Phi,\Phi^*)}$ is defined as in* (14), *then*

(i) $D_{(\Phi,\Phi^*)}(P, Q) \geq 0$ (*with equality only for* $P = Q$) *if and only if* Φ *is* (*strictly*) *convex*;
(ii) $D_{(\Phi,\Phi^*)}(P, Q) = 0$ *for all* $P, Q \in U$ *if and only if* Φ *is affine*.

We are now ready to relate functional Bregman divergences and score divergences. To that end, we extend the notion of a proper affine scoring rule to the set U, which may contain finite signed measures as $\mathcal{P} \subset U \subset \operatorname{span}\mathcal{P}$.

Definition 3.10. The affine scoring rule $\mathbf{S} : U \to \mathcal{A}(\mathcal{P})$ is said to be *proper* if $\mathbf{S}(P)(Q) \leq \mathbf{S}(Q)(Q)$ for all $P, Q \in U$, and *strictly proper*, if the latter inequality is strict for all $P \neq Q$.

A straightforward example of a proper affine scoring rule $\mathbf{S} : U \to \mathcal{A}(\mathcal{P})$ obtains by defining $\mathbf{S}$ as the affine component of a functional Bregman divergence, that is, by setting

$$\mathbf{S}(P)(Q) = \Phi^*(P) \cdot Q + \Phi(P) - \Phi^*(P) \cdot P,$$

for every $Q \in \operatorname{span}\mathcal{P}$, where Φ and Φ^* are the same as in Definition 3.8. As we show next, this actually constitutes a characterization of the proper affine scoring rules. The result first appeared in [17], Theorem 1, however, some modification is required to adapt it to the current measure-theoretic setting. Since the difference is only formal, we state the theorem without proof.

Theorem 3.11. *An affine scoring rule* $\mathbf{S} : U \to \mathcal{A}(\mathcal{P})$ *is* (*strictly*) *proper if and only if there is a* (*strictly*) *convex function* $\Phi : U \to \mathbb{R}$ *and a subgradient* $\Phi^* : U \to \mathcal{L}(\mathcal{P})$ *of* Φ *on* U *such that the expected score of* $\mathbf{S}$ *is in the form*

$$\mathbf{S}(P)(Q) = \Phi^*(P) \cdot (Q - P) + \Phi(P) \tag{15}$$

for all $P, Q \in U$.

In the theorem, the functional Bregman divergence associated with (Φ, Φ^*) is the score divergence of $\mathbf{S}$. For conditions describing whether and in what sense Φ and $\mathbf{S}$ may be uniquely associated, see Section 5. We next present a characterization of functional Bregman divergences by generalizing a result originally stated in the context of Euclidean spaces [3], Appendix A.

Theorem 3.12. *Let* $D : U \times U \to [0, \infty)$ *be a divergence on the convex set* $U \subset \operatorname{span}\mathcal{P}$. *Then* D *is a functional Bregman divergence on* U *if and only if*

(a) *for any* $P \in U$ *the function* $\Phi(Q) = D(P, Q)$ *is convex*;
(b) *there is a subgradient* $\Phi^* : U \to \mathcal{L}(\mathcal{P})$ *of* Φ *for which it holds that*

$$D(P, Q) = D_{(\Phi,\Phi^*)}(P, Q)$$

for all $P, Q \in U$.

Proof. Although it is not hard to adapt the proof from [3], Appendix A, we sketch it for completeness. The sufficient part of the claim is elementary. To show the necessary part, let D be the

Bregman divergence associated with the pair (Φ_1, Φ_1^*). Then

$$\Phi(Q) = D(P, Q) = \Phi_1(Q) - \Phi_1^*(P) \cdot Q + \Phi_1^*(P) \cdot P - \Phi_1(P)$$

and Φ_1 differ only by an affine function. Therefore, Φ is also convex and $\Phi^* : U \to \mathcal{L}(\mathcal{P})$ given by $\Phi^*(Q) = \Phi_1^*(Q) - \Phi_1^*(P)$ is its subgradient. It is now easy to see that the pairs (Φ, Φ^*) and (Φ_1, Φ_1^*) generate the same Bregman divergence. □

In conclusion, the last two results resolve the problem stated in [19], Section 2.2, to classify all score divergences. Indeed, this follows from Theorem 3.11, where score divergences are classified as functional Bregman divergences, which are in turn characterized by Theorem 3.12.

4. Characterization of proper scoring rules

Here we restrict attention to the important special case where scoring rules are defined over the set of probability measures $\mathcal{P}$, or its conic hull cone $\mathcal{P}$.

4.1. Direct characterization and order sensitivity

The analysis of Section 3 allow us to describe the classical proper scoring rules as follows: they are restrictions to the probability simplex of affine functionals whose graphs are supporting hyperplanes to a convex function – the entropy. Moreover, the score assigned to a particular outcome is the value at that outcome of the kernel function representing the associate affine functional. This may be seen more formally in the following result due to [19,32].

Theorem 4.1. *A scoring rule $S : \mathcal{P} \times \mathcal{X} \to \mathbb{R}$ is (strictly) proper if and only if there is a (strictly) convex function $\Phi : \mathcal{P} \to \mathbb{R}$ and a subgradient $\Phi^* : \mathcal{P} \to \mathcal{L}(\mathcal{P})$ of Φ such that*

$$S(P, x) = \Phi^*(P)(x) + \Phi(P) - \Phi^*(P) \cdot P,$$

for every $P \in \mathcal{P}$.

We next give a brief comment on the relationship between S, Φ, and Φ^* in the above theorem. Notice that every proper scoring rule S is a subgradient relative to $\mathcal{P}$ of its entropy with the additional property $S(P, P) = \Phi(P)$, for every $P \in \mathcal{P}$. Conversely, every subgradient Φ^* of Φ relative to $\mathcal{P}$ that satisfies $\Phi^*(P) \cdot P = \Phi(P)$ for every $P \in \mathcal{P}$ is a proper scoring rule on $\mathcal{P}$. We thus arrive at the following description of the type of subgradients that constitue a proper scoring rule.

Corollary 4.2. *Given a (strictly) convex function $\Phi : \mathcal{P} \to \mathbb{R}$ and a subgradient $\Phi^* : \mathcal{P} \to \mathcal{L}(\mathcal{P})$ of Φ, Φ^* is a (strictly) proper scoring rule with entropy Φ if and only if*

$$\Phi(P) = \Phi^*(P) \cdot P \tag{16}$$

for every $P \in \mathcal{P}$.

Unlike [19], Theorem 1, we do not allow scoring rules to become infinite, which is in line with the standard definition of subgradient in convex analysis. Moreover, subgradients may be prevented from becoming infinite if one introduces a suitable notion of interior for the domain of the associated entropy function.

Example 4.3. Let us consider the (negative) Shannon entropy,

$$\Phi(p) = \sum_{i=1}^{d} p_i \ln p_i, \qquad p \in \mathcal{P},$$

where $\mathcal{P}$ is the (relative) interior of the probability simplex in $\mathbb{R}^d$. It is easy to verify that Φ extends as a convex function to the positive orthant, $\mathbb{R}^d_+$. We have that

$$\nabla\Phi(p) = (\ln p_i + 1, \ldots, \ln p_d + 1), \qquad p \in \mathbb{R}^d_+,$$

is the gradient of Φ. Using Theorem 4.1, we find that the logarithmic scoring rule, $S_{\log}(p, i) = \ln p_i$, is a proper scoring rule associated with Shannon entropy.

We next present an alternative condition to properness. To that end, let $S : \mathcal{P} \times \mathcal{X} \to \mathbb{R}$ be a scoring rule and $P, Q \in \mathcal{P}$ be predictive distributions. For any $\lambda \in [0, 1]$, we set $P_\lambda = (1-\lambda)P + \lambda Q$ to be the convex mixture between P and Q.

Definition 4.4. We say that S is *(strictly) order sensitive* if the function

$$\lambda \to S(P_\lambda, Q), \qquad \lambda \in [0, 1],$$

is (strictly) monotone increasing for every $P, Q \in \mathcal{P}$.

Clearly, a scoring rule S that is (strictly) order sensitive is also (strictly) proper. In addition, the expected score of S improves whenever the original forecast P moves in the direction of the true density Q. Seemingly weaker, properness turns out to be equivalent to order sensitivity of a scoring rule.

Theorem 4.5 ([27]). *A scoring rule S is (strictly) proper if and only if S is (strictly) order sensitive.*

4.2. Characterization through 1-homogeneous entropies

We are next going to present another characterization of proper scoring rules which relies on Euler's homogeneous function theorem to incorporate condition (16) directly in the notion of subgradient. First, let us note that every function $\Phi : \mathcal{P} \to \mathbb{R}$ may be extended as a 1-homogeneous function to cone $\mathcal{P}$ by setting

$$\Phi(P) = (1 \cdot P)\Phi\left(\frac{P}{1 \cdot P}\right), \qquad \text{for every } P \in \operatorname{cone}\mathcal{P}.$$

Here, $1 \cdot P$ is the normalising constant of P. We recall that if Φ is convex on $\mathcal{P}$, then its 1-homogeneous extension to $\operatorname{cone}\mathcal{P}$ is a sublinear function. Similarly, every scoring rule $S : \mathcal{P} \times \mathcal{X} \to \mathbb{R}$ extends to a 0-homogeneous function on $\operatorname{cone}\mathcal{P}$ by setting

$$S(P, x) = S\left(\frac{P}{1 \cdot P}, x\right), \qquad \text{for every } P \in \operatorname{cone}\mathcal{P}.$$

A generalized version of *Euler's homogeneous function theorem* states that if $\Phi : \operatorname{cone}\mathcal{P} \to \mathbb{R}$ is 1-homogeneous, then

$$\partial\Phi(a) \cdot a = \Phi(a) \tag{17}$$

for every $a \in \operatorname{cone}\mathcal{P}$, where $\partial\Phi(a)$ is the subdifferential of Φ at a. The above identity relates sets, since $\partial\Phi(a)$ is generally a multi-valued map. It can be further shown that the subdifferential is a 0-homogeneous multi-valued map in the sense that it satisfies the relation $\partial\Phi(\lambda a) = \partial\Phi(a)$, for every $\lambda > 0$ and every $a \in \operatorname{cone}\mathcal{P}$ [20,29]. Consequently, it always possible to take a 0-homogeneous selection of subgradients to a 1-homogeneous function. Notice the similarity between conditions (16) and (17). In view of the latter, the definition of subgradient of a 1-homogeneous function simplifies as follows.

Definition 4.6. Let $\Phi : \operatorname{cone}\mathcal{P} \to \mathbb{R}$ be a 1-homogeneous function. Then $P^* \in \mathcal{L}(\mathcal{P})$ is a subgradient of Φ at P if for all $Q \in \operatorname{cone}\mathcal{P}$

$$\Phi(Q) \geq P^* \cdot Q, \tag{18}$$

with equality for $P = Q$. The subgradient P^* is called strict if the above inequality is strict for all P not positively collinear to Q.[10]

Clearly, the notion of subgradient to a 1-homogeneous function matches precisely the condition for properness of a scoring rule. Thus, we recover the classical characterisation of proper scoring rules due to [26] and [20]. We paraphrase the result by emphasizing the role of the extension of the entropy to the positive orthant for obtaining the "right" notion of subgradient, and simultaneously generalize the result to the context of probability measures. We also retain the original function-valued notation of scoring rules.

Theorem 4.7. *Let* $\mathbf{S} : \mathcal{P} \to \mathcal{L}(\mathcal{P})$ *be a scoring rule and* $\Phi : \mathcal{P} \to \mathbb{R}$ *be defined as* $\Phi(P) = \mathbf{S}(P) \cdot P$, *for every* $P \in \mathcal{P}$. *Then* $\mathbf{S}$ *is* (*strictly*) *proper if and only if the* 0*-homogeneous extension of* $\mathbf{S}$ *to* $\operatorname{cone}\mathcal{P}$ *is a* (*strict*) *subgradient of the* 1*-homogeneous extension of* Φ *to* $\operatorname{cone}\mathcal{P}$.

If we agree to call the 1-homogeneous extension of an entropy function a *1-homogeneous entropy*, then the theorem simply states that proper scoring rules are subgradients of 1-homogeneous entropies. Although Theorem 4.7 has only been proved in the context of probability densities by [20], the difference is purely notational and the proof remains unchanged.

[10] The convention for strict subgradient used in Definition 4.6 is standard in convex analysis in the context of 1-homogeneous functions. There is no danger of ambiguity, as we use it here only in the special context where Φ and S are defined on $\operatorname{cone}\mathcal{P}$ and Φ is 1-homogeneous.

In Section 2.4, we saw that the power scoring rule $\mathbf{S}_{\mathrm{pow}}(p)(q)$ is an affine function in $q \in \operatorname{cone}\mathcal{P}$. In contrast, the pseudospherical scoring rule $\mathbf{S}_{\mathrm{psu}}(p)(q)$ is linear in $q \in \operatorname{cone}\mathcal{P}$. In our next result, we characterize all such "linear" scoring rules.

Corollary 4.8. *Let* $\mathbf{S} : \operatorname{cone}\mathcal{P} \to \mathcal{A}(\mathcal{P})$ *be a proper affine scoring rule and let* $\Phi : \operatorname{cone}\mathcal{P} \to \mathbb{R}$, $\Phi(P) = \mathbf{S}(P)(P)$, *be the associated extended entropy. Then, the range of* $\mathbf{S}$ *lies in* $\mathcal{L}(\mathcal{P})$ *if and only if* $\mathbf{S}$ *is* 0*-homogeneous, if and only if* Φ *is* 1*-homogeneous.*

Proof. Suppose first that $\mathbf{S} : \operatorname{cone}\mathcal{P} \to \mathcal{L}(\mathcal{P})$. Then, in view of (15), $\mathbf{S} = \Phi^*$ is a subgradient of Φ that satisfies Euler's 1-homogeneous function identity,

$$\Phi^*(P) \cdot P = \Phi(P).$$

Hence, $\mathbf{S}$ is 0-homogeneous and Φ is 1-homogeneous. Conversely, if $\mathbf{S}$ is 0-homogeneous, or equivalently Φ is 1-homogeneous, it follows that the subgradients of Φ satisfy the above identity [20,29]. Then (15) reduces to $\mathbf{S} = \Phi^*$, implying that $\mathbf{S}$ is also a subgradient of Φ. □

4.3. Some examples

We continue with some illustrations of Theorem 4.7. We first do so in the context of discrete sample spaces.

Example 4.9. In contrast to Example 4.3, here we derive the logarithmic scoring rule from the 1-homogeneous extension to $\mathbb{R}^d_+$ of Shannon entropy,

$$\Phi(p) = \sum_{i=1}^{d} p_i \ln \frac{p_i}{|p|_1}, \qquad p \in \mathbb{R}^d_+,$$

where we have used the notation $|p|_1 = (p_1 + \cdots + p_d)$. In view of the fact that Φ is a differentiable function, we find that

$$\frac{\partial \Phi(p)}{\partial p_i} = \ln \frac{p_i}{|p|_1},$$

which is the 0-homogeneous extension of $S_{\log}$ to $\mathbb{R}^d_+$. The supporting hyperplanes of Φ become vertical at the boundary of $\mathbb{R}^d_+$, implying that the subdifferential of Φ is empty there. Consequently, the logarithmic scoring rule cannot be defined for probabilities that vanish.

We next present two examples in the context of continuous variables.

Example 4.10. Let us consider the Lebesgue space $L^2(\mathcal{X})$ and take $\mathcal{P}$ to be the set of all probability densities in $L^2(\mathcal{X})$ with respect to the Lebesgue measure on $\mathcal{X}$. We leave it to the reader to verify that the 1-homogeneous entropy,

$$\Phi(p) = \frac{\int_{\mathcal{X}} p^2(x)\,dx}{\int_{\mathcal{X}} p(x)\,dx}, \qquad p \in \operatorname{cone}\mathcal{P},$$

has a Gâteaux derivative on cone $\mathcal{P}$ given by

$$\Phi'(p)(x) = \frac{2p(x)}{\|p\|_1} - \left(\frac{\|p\|_2}{\|p\|_1}\right)^2, \qquad p \in \operatorname{cone}\mathcal{P}.$$

This immediately defines the proper scoring rule

$$S(p,x) = 2p(x) - \|p\|_2^2, \qquad p \in \mathcal{P},$$

known as the *quadratic scoring rule*. A more detailed and rigorous derivation of the quadratic scoring rule is also contained in Example 5.6.

Example 4.11. Consider the L^2-norm, $\Phi(p) = \|p\|_2$ in the Lebesgue space $L^2(\mathcal{X})$ and compute its Gâteaux derivative, $\Phi'(p)(x) = p(x)/\|p\|_2$, for $p \in \operatorname{cone}\mathcal{P}$. Since Φ is 1-homogeneous, this defines the proper scoring rule

$$S(p,x) = p(x)/\|p\|_2$$

known as the *spherical scoring rule*.

It was key in the above examples that we were able to extend the entropy Φ to cone $\mathcal{P}$, or span $\mathcal{P}$, where Φ was a differentiable function. See also [34] who studies proper scoring rules by making use of the duality theory of convex functions.

5. Regularity theory of entropy functions

Our main goal here is to understand under what conditions a general convex function constitutes the entropy of a proper scoring rule and whether the association between entropy and proper scoring rule is unique.

5.1. Entropy functions in finite dimensions

Below are summarized the most basic regularity properties of convex functions in Euclidean spaces, contained in any standard book on convex analysis such as [9,10,21,28,31].

Theorem 5.1. *Let $\phi : U \to \mathbb{R}$ be convex and $U \subset \mathbb{R}^n$ be an open convex set. Then*

(a) *ϕ is locally Lipschitz continuous on U and hence a continuous function*;
(b) *there is a dense subset A of U such that $U \setminus A$ has Lebesgue measure zero and $\nabla\phi$ exists on A*;
(c) *for every $a \in A$ we have*

$$\phi(x) \geq \nabla\phi(a) \cdot (x-a) + \phi(a), \qquad \text{for all } x \in U;$$

(d) *for every $a \in U \setminus A$ there is $a^* \in \mathbb{R}^n$ such that*

$$\phi(x) \geq a^* \cdot (x - a) + \phi(a), \qquad \text{for all } x \in U; \tag{19}$$

(e) *ϕ is differentiable everywhere on U if and only if $\nabla\phi$ is continuous on U if and only if ϕ is continuously differentiable on U.*

As a result, every convex function on an open convex set containing the probability simplex in $\mathbb{R}^d$ constitutes the entropy function of a proper (affine) scoring rule that is uniquely determined up to a negligible set from its entropy. Entropy functions are locally Lipschitz continuous and differentiable almost everywhere.

5.2. Entropy functions in infinite dimensions

The regularity theory of convex functions in normed spaces that are continuous is very similar to that in finite dimensions. The existence of discontinuous linear functionals in every infinite-dimensional normed space [9], Excercise 4.1.22, however, implies that continuity of convex functions is not a property that is preserved when passing to infinite dimensions. The continuous and discontinuous cases are qualitatively different and for that reason are considered separately in what follows. As prototypes to exemplify each of the two cases, we have the quadratic and Shannon entropy, respectively.

5.2.1. *Continuous case*

We assume we may identify the space of signed densities $\operatorname{span}\mathcal{P}$ with a normed space $\mathcal{N}$ and denote by $\mathcal{N}^*$ its topological dual. We further assume that the elements of $\mathcal{N}^*$ may be identified with elements of $\mathcal{L}(\mathcal{P})$, in particular they are real-valued functions on $\mathcal{X}$. We are interested in entropy functions of the form $\Phi : U \to \mathbb{R}$, where U is an open convex set in $\mathcal{N}$ containing $\mathcal{P}$. The following result gathers several equivalent characterizations of continuity of convex functions, see, for example, in [28], Section 3.5, or [9], Section 4.1.

Theorem 5.2. *Let U be an open convex set in a normed space $\mathcal{N}$, and let $\Phi : U \to \mathbb{R}$ be a convex function. Then the following assertions are equivalent*:

(i) *Φ is locally Lipschitz on U*;
(ii) *Φ is continuous on U*;
(iii) *Φ is continuous at some point of U*;
(iv) *Φ is locally bounded on U*;
(v) *Φ is bounded from above on a nonempty open subset of U.*

Thus, just like linear functionals, convex functions are either continuous at a single point or discontinuous and unbounded at every point of an open domain. When continuous, they are locally Lipschitz and thus "almost" differentiable in various technical senses which fall outside our scope [9], Section 4.6.

Subgradients in this context are defined as usual, but in addition we require them to be continuous linear functionals. In other words, any $p^* \in \mathcal{N}^*$ satisfying

$$\Phi(q) \geq p^* \cdot (q - p) + \Phi(p), \qquad \text{for all } q \in U; \tag{20}$$

is a subgradient of Φ at p. The set $\partial\Phi(p)$ of all such $p^* \in \mathcal{N}^*$ constitutes the *subdifferential* of Φ at p. The following fundamental result [28], Section A.3, called the supporting hyperplane theorem, Hahn–Banach separation theorem, or Mazur's theorem, guarantees the existence of subgradients of continuous convex functions.

Theorem 5.3. *Let U be an open convex set in a normed space $\mathcal{N}$, and let $\Phi : U \to \mathbb{R}$ be a continuous convex function. Then $\partial\Phi(a) \neq \varnothing$ for any $a \in U$.*

In order to describe when a continuous convex function has unique subgradients, we need a notion of derivative in normed spaces.

Definition 5.4. Let U be an open set in a normed space $\mathcal{N}$. A function $\Phi : U \to \mathbb{R}$ is *Gâteaux differentiable* at a point $p \in U$ if there is $p^* \in \mathcal{N}^*$ such that for every $q \in \mathcal{N}$ the limit

$$p^* \cdot q = \lim_{t \to 0} \frac{\Phi(p + tq) - \Phi(p)}{t} \tag{21}$$

exists. The functional p^* is called the *Gâteaux derivative* of Φ at p and is also denoted by $\Phi'(p)$.

In analogy to the Euclidean case, if Φ is continuous and the limit (21) exists for all $q \in \mathcal{N}$ and is linear in $q \in \mathcal{N}$, then the linear functional p^* it defines is necessarily a member of $\mathcal{N}^*$. The claim is a classic application of the Hahn–Banach theorem and may be used to justify the following result see, for example, [9], Section 4.2.

Theorem 5.5. *Let U be an open convex set in a normed space $\mathcal{N}$, and let $\Phi : U \to \mathbb{R}$ be a continuous convex function. Then Φ is Gâteaux differentiable at $a \in U$ if and only if $\partial\Phi(a)$ is a singleton. In this case, $\partial\Phi(a) = \{\Phi'(a)\}$.*

We finish this subsection with the following rigorous derivation of the power scoring rule.

Example 5.6. Let us consider the Lebesgue space $\mathcal{N} = L^\gamma(\mathcal{X})$, for $1 < \gamma < \infty$, and take $\mathcal{P}$ to be the set of all probability densities in $\mathcal{N}$ with respect to the Lebesgue measure on $\mathcal{X}$. The continuous dual space is $\mathcal{N}^* = L^{\gamma/(\gamma-1)}(\mathcal{X})$. Consider the *power* or *Tsalis* entropy function,

$$\Phi_\gamma(p) = \int_{\mathcal{X}} \left|p(x)\right|^\gamma dx,$$

where $p \in \mathcal{N}$ is arbitrary. We proceed to compute the Gâteaux derivative of Φ_γ on $L^\gamma(\mathcal{X})$. For $p, q \in L^\gamma(\mathcal{X})$ and $p_t = p + tq$, we have

$$\begin{aligned} \lim_{t \to 0} \frac{\Phi_\gamma(p + tq) - \Phi_\gamma(p)}{t} &= \frac{d}{dt}\bigg|_{t=0} \int_{\mathcal{X}} \left|p_t(x)\right|^\gamma dx \\ &= \gamma \operatorname{sgn}(p)|p|^{\gamma-1} \cdot q. \end{aligned}$$

The exchange of differentiation and integration is legitimate due to the fact that the resulting integrand is a Lebesgue integrable function. In fact, $\Phi'_\gamma(p)(x) = \gamma \operatorname{sgn}(p(x))|p(x)|^{\gamma-1} \in \mathcal{N}^*$ and thus Φ'_γ is indeed the Gâteaux derivative of Φ_γ. We therefore may write that

$$D_\gamma(p,q) = \|q\|_\gamma^\gamma - \gamma \operatorname{sgn}(p)|p|^{\gamma-1} \cdot (q-p) - \|p\|_\gamma^\gamma$$

is the Bregman divergence on $L^\gamma(\mathcal{X})$ associated with Φ_γ. On $\mathcal{P}$, the associated proper scoring rule is

$$S_\gamma(p,x) = \gamma p^{\gamma-1}(x) - (\gamma-1)\|p\|_\gamma^\gamma,$$

the *power scoring rule with exponent* γ.

5.2.2. *Discontinuous case*

The most basic regularity property common to all convex functions is the existence of directional derivatives. Through directional derivatives one may connect the finite and infinite-dimensional regularity theories of convex functions. Throughout this subsection, the set U is convex and we have that $\mathcal{P} \subset U \subset \operatorname{span}\mathcal{P}$. The elements of $\mathcal{P}$ are considered to be represented by their probability densities.

Definition 5.7. The *right directional derivative* of $\Phi : U \to \mathbb{R}$ at $p \in U$ along the vector $q \in \operatorname{cone}(U-p)$ is defined as the limit

$$\Phi'_+(p;q) = \lim_{t\to 0^+} \frac{\Phi(p+tq) - \Phi(p)}{t}, \tag{22}$$

whenever it exists.

Some authors prefer to write the limit (22) in the following equivalent way

$$\Phi'_+(p;q-p) = \lim_{t\to 0^+} \frac{\Phi((1-t)p+tq) - \Phi(p)}{t},$$

for all $p, q \in U$. Equivalency follows from the identity

$$\operatorname{cone}(U-p) = \{\lambda(q-p) | q \in U, \lambda > 0\}$$

and the observation that the latter gives all directions based at p towards points in U. It is well known (from the finite-dimensional theory) that when Φ is convex, we have that

$$\Phi'_+(p;q) = \inf_{t>0} \frac{\Phi(p+tq) - \Phi(p)}{t}.$$

Hence, allowing convergence to $-\infty$, the limit always exists. Moreover, the limit is finite, provided that p lies in the relative interior of a line segment in U along the direction q [31], Theorem 24.1.

As a way of informally justifying what follows, let us restrict Φ to some finite dimensional open convex subset U' of U. Then the graph of $q \to \Phi'_+(p; q-p)$, where p is fixed and $q \in U'$ is arbitrary, is a convex cone[11] with vertex at p that supports the graph of Φ at p and is the envelope of all supporting hyperplanes of Φ at p. Similar arguments lead to the following proposition, see e.g. the proof of [28], Lemma 1.5.1, [29], Proposition 2.3, or [31], Theorem 23.1.

Proposition 5.8. *Suppose that $\Phi : U \to \mathbb{R}$ is convex. Then*

$$\Phi(q) \geq \Phi'_+(p; q-p) + \Phi(p) \qquad \text{for all } p, q \in U$$

and the function $\Phi'_+(p; \cdot) : \operatorname{cone}(U-p) \to \mathbb{R} \cup \{-\infty\}$ is sublinear.

To help understand the following theorem, let us assume that $\Phi'_+(p; \cdot) : \operatorname{cone}(U-p) \to \mathbb{R}$ is finite, for some $p \in U$. Then, starting with any subgradient of Φ at p relative to U', we may use the Hahn–Banach theorem[12] to extend this subgradient to a linear functional $l_p : \operatorname{span}\mathcal{P} \to \mathbb{R}$ such that

$$l_p(q) \leq \Phi'_+(p; q)$$

for all $q \in \operatorname{cone}(U-p)$. Moreover, $\Phi'_+(p; q)$ will be the envelope of all such l_p. Due to the fact that Φ is not continuous, we cannot guarantee that the resulting l_p would be bounded with respect to some norm. Instead, we may only verify in each concrete case whether l_p identifies with a member p^* of $\mathcal{L}(\mathcal{P})$, which, if so, will constitute a subgradient in our framework. Thus, we arrive at the following result, which generalizes [29], Theorem 3.1, where Φ is assumed to be a 1-homogeneous convex function on $\operatorname{cone}\mathcal{P}$.

Theorem 5.9. *A convex function $\Phi : U \to \mathbb{R}$ has a subgradient $p^* \in \mathcal{L}(\mathcal{P})$ at $p \in U$ if and only if*

$$p^* \cdot q \leq \Phi'_+(p; q)$$

for all $q \in \operatorname{cone}(U-p)$.

We note that in the above result if there is a single $q \in \operatorname{cone}(U-p)$ such that $\Phi'_+(p; q) = -\infty$, then no $p^* \in \mathcal{L}(\mathcal{P})$ exists that satisfies the above inequality and thus $\partial\Phi(p)$ is empty. The condition that $\Phi'_+(p; q)$ is finite for all $q \in \operatorname{cone}(U-p)$ cannot be guaranteed for a general convex function Φ. It must hold, however, whenever Φ is the entropy function of a proper scoring rule. The latter crucially depends on the suitable choice of domain U and ambient space $\operatorname{span}\mathcal{P}$ with respect to Φ.

The last question we consider here is that of uniqueness of subgradients. Most entropy functions of practical significance are defined as integral functionals of smooth convex kernels. We thus typically have that $\Phi'_+(p; q) = p^* \cdot q$, for some $p^* \in \mathcal{L}(\mathcal{P})$, where $\mathcal{P}$ and U are suitably

[11] The cone may be flat, i.e. a hyperplane, if Φ is smooth in a neighbourhood of p.

[12] In fact, we need a slight modification of the classical theorem to allow for sublinear functions that are finite only on a convex subcone of the ambient vector space, see e.g. [29], Theorem B.4.

chosen, and $q \in \operatorname{cone}(U - p)$ is arbitrary. Notice that $\operatorname{cone}(U - p)$ is generally not a pointed cone and may contain some straight lines through the origin. Let $\mathcal{O}(p)$ denote the maximal linear subspace of $\operatorname{cone}(U - p)$, or equivalently, the set of all vectors $q \in \operatorname{cone}(U - p)$ such that $-q$ is in $\operatorname{cone}(U - p)$ too. The latter may also be written as

$$\mathcal{O}(p) = \operatorname{cone}(U - p) \cap -\operatorname{cone}(U - p).$$

Let us assume that the directional derivative $\Phi'_+(p; \cdot)$ is linear on $\mathcal{O}(p)$. Then, any subgradient p^* of Φ at p is uniquely defined in all directions in $\mathcal{O}(p)$ as a result of finite-dimensional theory. The uniqueness of p^* as a functional on $\operatorname{span}\mathcal{P}$ is therefore dependent on the size of the subspace $\mathcal{O}(p)$ of $\operatorname{span}\mathcal{P}$. More precisely, p^* is uniquely defined as an element of $\mathcal{L}(\mathcal{P})$ if and only if the space $\mathcal{O}(p)$ has trivial annihilator in $\mathcal{L}(\mathcal{P})$. To clarify the latter notion, let us recall that if E is a subset of $\operatorname{span}\mathcal{P}$, the linear subspace of $\mathcal{L}(\mathcal{P})$ of all f such that $f \cdot p = 0$ is the *annihilator* of E in $\mathcal{L}(\mathcal{P})$. The annihilator is called *trivial* whenever it is identical to the trivial vector space $\{0\}$. We next give a crude algebraic analogue of the topological notion of *quasi-interior* [7], Definition 2.6 or [8], Definition 2.3.

Definition 5.10. Any point $p \in U$ such that $\mathcal{O}(p)$ has trivial annihilator in $\mathcal{L}(\mathcal{P})$ is called an *algebraically quasi-interior* point of U relative to $\operatorname{span}\mathcal{P}$. The collection of all algebraically quasi-interior points of U is the *algebraic quasi-interior* of U, denoted by $\operatorname{aqi} U$.

As an illustration, notice that the algebraic quasi-interior coincides with the interior of U with respect to $\operatorname{span}\mathcal{P}$, whenever the latter is finite dimensional. A convex set in an infinite-dimensional normed space, however, may have empty interior but nonempty algebraic quasi-interior. We also have that for any two points p_1 and p_2 in $\operatorname{aqi} U$ the line segment connecting them also lies in $\operatorname{aqi} U$. Indeed, consider $p_\lambda = (1 - \lambda)p_1 + \lambda p_2$, for $\lambda \in (0, 1)$, and notice that any line segment in U containing p_1 in its relative interior has a homothetic copy containing p_λ in its relative interior that also lies in U. It follows that $\mathcal{O}(p_\lambda) \supset \mathcal{O}(p_1)$ and $\mathcal{O}(p_\lambda) \supset \mathcal{O}(p_2)$. Thus, the annihilator of $\mathcal{O}(p_\lambda)$ must be a subspace of the intersection of the annihilator of $\mathcal{O}(p_1)$ and the annihilator of $\mathcal{O}(p_2)$. We conclude that $\operatorname{aqi} U$ is a convex subset of U. With the arguments directly preceding the above definition, and relying on Proposition 5.8, we have shown the following result.

Theorem 5.11. *Consider a convex function $\Phi : U \to \mathbb{R}$ and a point $p \in \operatorname{aqi} U$. If there is $p^* \in \mathcal{L}(\mathcal{P})$ such that*

$$p^* \cdot q = \Phi'_+(p; q) \tag{23}$$

for all $q \in \operatorname{cone}(U - p)$, then p^* is the unique subgradient of Φ at p that belongs to $\mathcal{L}(\mathcal{P})$.

In the special case of $U = \operatorname{cone}\mathcal{P}$ and Φ being 1-homogeneous and convex, this result has been proven in [29], Theorem 3.2. We use it to construct positive cones of densities with nonempty algebraic quasi-interior where the logarithmic and Hyvärinen scoring rules are the unique 0-homogeneous subgradients of their 1-homogeneous entropy functions [29], Section 4.

Acknowledgement

Part of this work has been accomplished while at Heidelberg Institute for Theoretical Studies, Germany.

References

[1] Abernethy, J.D. and Frongillo, R.M. (2012). A characterization of scoring rules for linear properties. *JMLR Workshop and Conference Proceedings*: *COLT* **23** 27.1–27.13.
[2] Banerjee, A., Guo, X. and Wang, H. (2005). On the optimality of conditional expectation as a Bregman predictor. *IEEE Trans. Inform. Theory* **51** 2664–2669. MR2246384
[3] Banerjee, A., Merugu, S., Dhillon, I.S. and Ghosh, J. (2005). Clustering with Bregman divergences. *J. Mach. Learn. Res.* **6** 1705–1749.
[4] 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.
[5] Bauschke, H.H. and Borwein, J.M. (2001). Joint and separate convexity of the Bregman distance. In *Inherently Parallel Algorithms in Feasibility and Optimization and Their Applications* (*Haifa*, 2000). *Stud. Comput. Math.* **8** 23–36. Amsterdam: North-Holland. MR1853214
[6] Boissonnat, J.-D., Nielsen, F. and Nock, R. (2010). Bregman Voronoi diagrams. *Discrete Comput. Geom.* **44** 281–307. MR2671013
[7] Borwein, J. and Goebel, R. (2003). Notions of relative interior in Banach spaces. *J. Math. Sci.* (*N. Y.*) **115** 2542–2553. MR1992991
[8] Borwein, J.M. and Lewis, A.S. (1992). Partially finite convex programming, Part I: Quasi relative interiors and duality theory. *Math. Program.* **57** 15–48.
[9] Borwein, J.M. and Vanderwerff, J.D. (2010). *Convex Functions*: *Constructions*, *Characterizations and Counterexamples*. *Encyclopedia of Mathematics and Its Applications* **109**. Cambridge: Cambridge Univ. Press. MR2596822
[10] Boyd, S.P. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge: Cambridge Univ. Press.
[11] Bröcker, J. (2009). Reliability, sufficiency, and the decomposition of proper scores. *Q. J. R. Meteorol. Soc.* **135** 1512–1519.
[12] Dawid, A.P. and Musio, M. (2014). Theory and applications of proper scoring rules. *Metron* **72** 169–183. MR3233147
[13] de Finetti, B. (1975). *Theory of Probability*: *A Critical Introductory Treatment*. *Vol.* 2. London: Wiley. MR0440641
[14] Fissler, T. and Ziegel, J.F. (2015). Higher order elicitability and Osband's principle. Preprint.
[15] Frigyik, B.A., Srivastava, S. and Gupta, M.R. (2008). Functional Bregman divergence and Bayesian estimation of distributions. *IEEE Trans. Inform. Theory* **54** 5130–5139. MR2589887
[16] Frongillo, R.M. (2013). Eliciting Private Information from Selfush Agents, Ph.D. thesis, University of California, Berkeley.
[17] Frongillo, R.M. and Kash, I. (2014). General truthfulness characterizations via convex analysis. In *Web and Internet Economics*. *Lecture Notes in Computer Science* **8877** 354–370. Springer.
[18] Fujisawa, H. and Eguchi, S. (2008). Robust parameter estimation with a small bias against heavy contamination. *J. Multivariate Anal.* **99** 2053–2081. MR2466551
[19] Gneiting, T. and Raftery, A.E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* **102** 359–378. MR2345548
[20] Hendrickson, A.D. and Buehler, R.J. (1971). Proper scores for probability forecasters. *Ann. Math. Statist.* **42** 1916–1921. MR0314430

[21] Hiriart-Urruty, J.-B. and Lemaréchal, C. (2001). *Fundamentals of Convex Analysis*. *Grundlehren Text Editions*. Berlin: Springer. MR1865628
[22] Kanamori, T. and Fujisawa, H. (2014). Affine invariant divergences associated with proper composite scoring rules and their applications. *Bernoulli* **20** 2278–2304. MR3263105
[23] Kanamori, T. and Fujisawa, H. (2015). Robust estimation under heavy contamination using enlarged models. *Biometrika* **102** 559–572.
[24] Lambert, N. (2013). Elicitation and evaluation of statistical forecasts. Working paper.
[25] Lindley, D.V. (1982). Scoring rules and the inevitability of probability. *Int. Stat. Rev.* **50** 1–26. MR0668607
[26] McCarthy, J. (1956). Measures of the value of information. *Proc. Natl. Acad. Sci. USA* **42** 654–655.
[27] Nau, R.F. (1985). Should scoring rules be 'effective'? *Manage. Sci.* **31** 527–535.
[28] Niculescu, C.P. and Persson, L.-E. (2006). *Convex Functions and Their Applications*: *A Contemporary Approach*. *CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC* **23**. New York: Springer. MR2178902
[29] Ovcharov, E.Y. (2015). Existence and uniqueness of proper scoring rules. *J. Mach. Learn. Res.* **16** 2207–2230. MR3450505
[30] Predd, J.B., Seiringer, R., Lieb, E.H., Osherson, D.N., Poor, H.V. and Kulkarni, S.R. (2009). Probabilistic coherence and proper scoring rules. *IEEE Trans. Inform. Theory* **55** 4786–4792. MR2597577
[31] Rockafellar, R.T. (1972). *Convex Analysis*, 2nd ed. *Princeton Mathematical Series*. Princeton: Princeton Univ. Press.
[32] Savage, L.J. (1971). Elicitation of personal probabilities and expectations. *J. Amer. Statist. Assoc.* **66** 783–801.
[33] Steinwart, I., Pasin, C., Williamson, R.C. and Zhang, S. (2014). Elicitation and identification of properties. *JMLR Workshop and Conference Proceedings*: *COLT* **35** 482–526.
[34] Williamson, R.C. (2014). The geometry of losses. *JMLR Workshop and Conference Proceedings*: *COLT* **35** 1078–1108.
[35] Zălinescu, C. (2002). *Convex Analysis in General Vector Spaces*. River Edge, NJ: World Scientific Co. MR1921556

Received August 2015 and revised March 2016