Proper Scoring Rules and Bregman Divergences

Evgeni Y. Ovcharov *
Heidelberg Institute for Theoretical Studies
Schloss-Wolfsbrunnenweg 35, D-69118 Heidelberg, Germany
tru1r6@yahoo.com

February 5, 2015

Abstract

We present a new perspective on proper scoring rules (PSRs) in which they are naturally derived from a more general convex construction that also includes functional Bregman divergences. Our motivation lies in the fact that sets of probability distributions are negligible, mathematically, and functions defined exclusively on such sets do not have good analytical properties. We first examine the fact that all entropy functions may be extended as sublinear functions of denormalised probability densities and PSRs are characterised as their subgradients. We then proceed to explore general convex extensions of entropy functions and uncover the connection between functional Bregman divergences and PSRs. We examine and systematise previous characterisation results of PSRs and Bregman divergences in a unified theoretical framework.

Keywords: Proper scoring rules, Bregman divergence, Extensions, Characterisation, Regularity, Functional derivative, Subgradient, Convex.

1 Introduction

The theory of PSRs originated in probabilistic forecasting as a way to evaluate the quality of probabilistic predictions and to motivate a forecaster to be honest in the probabilities he announces (Brier 1950; Good 1952; Savage 1971). PSRs are used in disparate fields such as meteorology, finance, and pattern classification by an algorithm or a forecaster who attempts to optimise the average score to yield refined, calibrated probabilities (Gneiting and Katzfuss 2014; Bröcker 2009; Dawid and Musio 2014). Our main concern here is the mathematical theory of PSRs.

*The author has been supported by the European Union Seventh Framework Programme under grant agreement no. 290976

AMS 2000 subject classifications: Primary 62C99; Secondary 62A99, 26B25
In measure and topology, a set of probability distributions $\mathcal{P}$ is negligible since it is a subset of the unit simplex, which lies on the surface of the unit sphere in the 1-norm. As a result of this, we may not expect good analytical properties from quantities defined solely on such sets. A natural candidate for an extended domain is the conic hull of $\mathcal{P}$, denoted cone $\mathcal{P}$. This set is statistically relevant and arises in connection to denormalised probability models. It is well-known that every entropy function has a sublinear extension to cone $\mathcal{P}$, which allows PSRs to be characterised as subgradients of sublinear functions (Hendrickson and Buehler 1971; Ovcharov 2014). For its theoretical importance, this extension is referred by us as canonical.

Other convex extensions of entropy functions are also fundamentally significant. In this more general context, we may identify the expected scores of PSRs as supporting hyperplanes to the graphs of the associated entropy functions. From a slightly different perspective, Frongillo and Kash 2014 also arrive at the same concept and term the expected scores as affine scores. The associated divergences form the class of functional Bregman divergences. In contrast to earlier works such as Frigyik, Srivastava, and Gupta 2008, we study here functional Bregman divergences under natural regularity conditions for PSRs.

In the present paper, we present a new perspective on PSRs in which they are naturally derived from a more general convex construction that also includes functional Bregman divergences. In Section 2, we set the stage by introducing the canonical extension of entropy functions and PSRs and examine its properties. In Section 3, we proceed to explore general convex extensions of entropy functions and bring to light the connection between functional Bregman divergences and PSRs. In a separate subsection we present many of the important properties of these divergences, which we have adapted or extended from more specific contexts. Finally, in Section 4, we describe all PSRs associated with a given entropy function. To that end, we present some results that characterise the subdifferentials of entropy functions. We accompany our key points with examples and discussions throughout the text.

1.1 Some motivational examples

Many statistical models contain unknown normalising constants whose computation is too difficult to be done in practice. In the method of score matching, one applies the Hyvärinen scoring rule directly to denormalised probability densities to estimate their parameters (Dawid and Musio 2012; Dawid and Musio 2014; Hyvärinen 2005; Hyvärinen 2007). In robust estimation, one also deals with denormalised models. The reason here is because only a certain part (referred to as the target ratio) of the sample is drawn from the true or target density. Kanamori and Fujisawa 2014 show that the power scoring rules may be employed to estimate both the target ratio and the target density. To that end, they minimise the extended divergence of power scoring rules to estimate the parameters of a denormalised statistical model. It is also worth noting that PSRs may be regarded as a special type of allocation rules in a class of Bayesian games, which too, in the general case, are defined on broader domains (Frongillo
and Kash 2014). And lastly, there is an interesting connection to the problem of elicitation of point-forecasts in terms of PSRs (Steinwart et al. 2014; Lambert 2013; Ziegel 2014), where the consistent scoring functions for linear properties, such as expectation and higher moments, are characterised as Bregman divergences (Abernethy and Frongillo 2012).

In Ovcharov 2014, we consider the question in what sense a scoring rule may be regarded to be uniquely associated with its entropy function Φ. Geometrically, we must extend Φ to a sufficiently large domain where Φ has a unique supporting hyperplane at every point on its domain except perhaps on a negligible set of points. What constitutes a negligible set depends on the context but typical quantifiers are measure, dimension, and Baire category. Under the appropriate notion, the domain of Φ must itself not be a negligible set. Natural extended domains are the conic and linear hulls of \( \mathcal{P} \), and other sufficiently large convex sets. For example, the power scoring rules and the pseudospherical scoring rules have differentiable entropy functions after suitable extensions to Lebesgue spaces.

An interesting problem arises from the fact that many common function spaces, e.g. the Lebesgue spaces on \( \mathbb{R}^d \), have positive cones (the set of nonnegative functions) with empty interior (Borwein and Lewis 1992; Ovcharov 2014). In this context, for example, Shannon and Hyvärinen entropy functions cannot be naturally defined for densities that change sign and therefore have domains with empty interior. (This remains generally true for the entropy functions of proper local scoring rules of higher orders.) Although one cannot define a gradient for these entropies, their sublinear extensions have unique supporting hyperplanes on dense subsets of their extended domains and uniquely associated PSRs (Ovcharov 2014). The result shows again the advantage in regarding PSRs natively as restrictions of more general quantities.

### 2 The canonical extension

We begin with some formal definitions. In our setup, we consider the distributions in the family \( \mathcal{P} \) to be continuous with respect to a fixed measure \( \mu \) on \( \mathcal{X} \) and represented by their probability densities. We call the functions \( f : \mathcal{X} \rightarrow \mathbb{R} \) \( \mathcal{P} \)-integrable if

\[
\int_{\mathcal{X}} |f(x)| p(x) d\mu(x) < \infty
\]

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

For each predictive density \( q \in \mathcal{P} \) and each realisation \( x \in \mathcal{X} \), a scoring rule \( S \) assigns a numerical score \( S(q)(x) \). Formally, a scoring rule \( S \) is a mapping \( S : \mathcal{P} \rightarrow \mathcal{L}(\mathcal{P}) \). This definition allows us to regard \( S(q)(X) \) as a random function of \( X \) with finite expectation. The condition reflects the fact that the true density of \( X \) is generally unknown and potentially any \( p \in \mathcal{P} \). The expectation of \( S(q) \) under the true density \( p \) is denoted by

\[
p \cdot S(q) := \int_{\mathcal{X}} S(q)(x)p(x)d\mu(x)
\]
and also termed the expected score of $S$.

A scoring rule $S$ that maximises its expected score at the true density,

$$p \cdot S(p) = \max_{q \in \mathcal{P}} p \cdot S(q),$$

is called proper. If the true density is always a unique maximiser, we call $S$ strictly proper. The function

$$\Phi(q) = p \cdot S(p),$$

for every $p \in \mathcal{P}$, is convex as a pointwise maximum of linear functions. We refer to it as the (negative) entropy function associated with $S$.

It is key for us to find a notion of subgradient suitable for PSRs. Generally, subgradients are linear functionals on the linear span of $\mathcal{P}$, denoted $\text{span} \mathcal{P}$, and form the algebraic dual space of $\mathcal{P}$. In the present context, we additionally require that subgradients are members of the space $\mathcal{L}(\mathcal{P})$, which may be naturally identified with a subspace of the algebraic dual space of $\mathcal{P}$. In special cases, we may further introduce a topology on $\mathcal{P}$ and require that subgradients are continuous with respect to that topology. In this section, however, we shall only be concerned with subgradients in the more general algebraic setting.

**Definition 2.1.** Let $\mathcal{K}$ be a convex set in $\text{span} \mathcal{P}$ containing $\mathcal{P}$. Given a function $\Phi : \mathcal{K} \to \mathbb{R}$ and a point $q \in \mathcal{K}$, we say that $q^* \in \mathcal{L}(\mathcal{P})$ is a (\{$\mathcal{P}$\}-integrable) subgradient of $\Phi$ at $q$ relative to $\mathcal{K}$ if

$$\Phi(p) \geq (p - q) \cdot q^* + \Phi(q)$$

for all $p \in \mathcal{K}$. If the above inequality is strict for all $p \neq q$, the subgradient $q^*$ is called strict.

Geometrically, the affine functional in (3), $p \to (p - q) \cdot q^* + \Phi(q)$, defines a supporting hyperplane to $\Phi$ at $q$. The subgradient $q^*$ is a gradient to that supporting hyperplane in the setting of Euclidean spaces and generalises that notion in infinite-dimensional spaces. There is a one-to-one correspondence between subgradients and supporting hyperplanes at a given point. The collection of all subgradients of $\Phi$ at $q$ is called the subdifferential of $\Phi$ at $q$ and denoted by $\partial \Phi(q)$. Under our regularity assumptions, $\partial \Phi(q)$ is a subset of $\mathcal{L}(\mathcal{P})$. We characterise the subdifferentials of an entropy function $\Phi$ in Section 4. Suppose now that $\partial \Phi(q) \neq \emptyset$ for each $q \in \mathcal{K}$. Then, we call a selection of subgradients $S(q) \in \partial \Phi(q)$, for each $q \in \mathcal{K}$, a subgradient of $\Phi$ on $\mathcal{K}$.

In view of (1) and (2), every PSR $S$ is a subgradient of its entropy function $\Phi$ on $\mathcal{P}$. The converse is also true if $S$ satisfies $q \cdot S(q) = \Phi(q)$, for every $q \in \mathcal{P}$. This suggests a connection with Euler’s homogeneous function theorem which we exploit in what follows. To that end, let us review some properties related to homogeneity. First, for two sets $A$ and $B$ in $\text{span} \mathcal{P}$, we employ the Minkowski sum and difference notation: $A \pm B = \{a \pm b \mid a \in A, b \in B\}$. For $\lambda \in \mathbb{R}$ and $A \subseteq \text{span} \mathcal{P}$, we set $\lambda A = \{\lambda a \mid a \in A\}$. A set $C \subseteq \text{span} \mathcal{P}$ is called a convex cone
if $\lambda C = C$ and $C + C = C$ for all $\lambda > 0$. Throughout, we take the conical hull of a set $C$, denoted $\text{cone } C$, to mean the smallest convex cone that contains $C$. Note that this differs from another common definition of a conical hull which also adds the origin. Let a function $f : C \to \mathbb{R}$ be given, where $C$ is a convex cone. It is said that $f$ is $\alpha$-homogeneous for some $\alpha \in \mathbb{R}$ if $f(\lambda q) = \lambda^\alpha f(q)$ for every $q \in C$ and every $\lambda > 0$. An extended version of Euler’s homogeneous function theorem states that if $\Phi : C \to \mathbb{R}$ is 1-homogeneous, then

$$q \cdot \partial \Phi(q) = \Phi(q)$$

for every $q \in C$ (Hendrickson and Buehler 1971; Ovcharov 2014). In the above identity, $\partial \Phi(q)$ is a multi-valued map, so the identity must be interpreted to hold for each element separately. It can be shown further that the subdifferential is a 0-homogeneous multi-valued map in a sense that it satisfies the relation $\partial \Phi(\lambda q) = \partial \Phi(q)$, for every $\lambda > 0$ and every $q \in C$. (For example, combine Ovcharov 2014, Proposition 2.3 (c) with Theorem 4.2 below for easy proof.)

The most important convex cone in the present context is the set $\text{cone } P = \{\lambda p | \lambda > 0, p \in P\}$. Given a scoring rule $S : P \to \mathcal{L}(P)$, we may extend $S$ to $\text{cone } P$ as a 0-homogeneous mapping by setting

$$S(q) = S\left(\frac{q}{q \cdot 1}\right),$$

for every $q \in \text{cone } P$, where $q \cdot 1$ is the normalising constant of $q$. For any $\Phi : P \to \mathbb{R}$, the 1-homogeneous extension of $\Phi$ is given by

$$\Phi(q) = (q \cdot 1)\Phi\left(\frac{q}{q \cdot 1}\right)$$

for every $q \in \text{cone } P$. Notice that if $\Phi(p) = p \cdot S(p)$, for every $p \in \text{cone } P$, and $S$ is 0-homogeneous, then $\Phi$ is automatically 1-homogeneous. Due to (4), in the context of 1-homogeneous functions, Definition 2.1 reduces to the following.

**Definition 2.2.** Given a 1-homogeneous function $\Phi : \text{cone } P \to \mathbb{R}$ and a point $q \in \text{cone } P$, we say that $q^* \in \mathcal{L}(P)$ is a ($P$-integrable) subgradient of $\Phi$ at $q$ relative to $\text{cone } P$ if

$$\Phi(p) \geq p \cdot q^*$$

for all $p \in \text{cone } P$, with equality for $p = q$. If the above inequality is strict for all $p$ not positively collinear to $q$, the subgradient $q^*$ is called strict.

In Definition 2.2, we adopt a special convention about strict convexity. Notice that with the usual definition $\Phi$ cannot be strictly convex because the identity $\Phi(\lambda q) = \lambda \Phi(q)$, for every $\lambda > 0$, implies that $\Phi(q + q) = \Phi(q) + \Phi(q)$. Therefore, it is reasonable to call a 1-homogeneous function $\Phi$ strictly convex if the restriction of $\Phi$ to $P$ is a strictly convex function. The same applies for strict subgradients, which must satisfy (5) with strict inequalities only for $p$ and $q$ that are not positively collinear. Moreover, every convex 1-homogeneous function $\Phi$
is a sublinear function. In the same vein as above, we call $\Phi$ strictly\ sublinear if its restriction to $P$ is strictly convex. In view of Definition 2.2, any subgradient $S$ on cone $P$ of a 1-homogeneous function $\Phi$ satisfies (1), and hence $S$ is a PSR. Strict propriety is equivalent to the condition that for any two $p, q \in \text{cone } P$ that are not positively collinear, $\partial \Phi(p) \cap \partial \Phi(q) = \emptyset$, and thus it is equivalent to strict convexity of $\Phi$. Thus we recover the classical results of McCarthy 1956 and Hendrickson and Buehler 1971. In our formulation, we emphasise the connection between propriety and sublinear extensions of the entropy.

**Theorem 2.3.** Let $S : P \to \mathcal{L}(P)$ be a scoring rule and $\Phi : P \to \mathbb{R}$ be defined as $\Phi(p) = p \cdot S(p)$, for every $p \in P$. Then $S$ is (strictly) proper if and only if the 0-homogeneous extension of $S$ is a (strict) subgradient of the 1-homogeneous extension of $\Phi$ on cone $P$.

We remark that Theorem 2.3 employs a stronger notion of $P$-integrability than the analogous result by Hendrickson and Buehler 1971. Specifically, we require absolute convergence of the integrals, which is the standard notion of convergence in Lebesgue integration, while in the original version of the theorem only relative convergence is assumed. Our notion of $P$-integrability makes it possible in many situations to identify the spaces $\mathcal{L}(P)$ as subspaces of standard function spaces. (See Example 4.3 below or Ovcharov 2014, Section 4.)

Williamson 2014 gives a different perspective on Theorem 2.3 relying on duality theory of convex functions. Due to the theoretical importance of the extension of PSRs and entropy functions described in Theorem 2.3 and the text above it, we call this extension canonical. The most basic implication of it is the fact that PSRs may be identified as subgradients of sublinear functions. We discuss other extensions in Section 3.

Let us now consider an alternative characterisation of PSRs due to Gneiting and Raftery 2007.

**Theorem 2.4.** A scoring rule $S : P \to \mathcal{L}(P)$ is (strictly) proper if and only if there exists a (strictly) convex function $\Phi : P \to \mathbb{R}$ such that

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

(6)

for every $q \in P$, where $\Phi^*(q)$ is a subgradient of $\Phi$ at $q$ with respect to $P$.

Notice the apparent lack of symmetry in the above theorem. On the one hand, every PSR is a subgradient on $P$ of its entropy function. On the other hand, if $\Phi^*$ is an arbitrary subgradient of $\Phi$ on $P$, then one needs to add an additional term, $\Phi(q) - q \cdot \Phi^*(q)$, to $\Phi^*(q)$ to obtain a subgradient that is a PSR associated with $\Phi$. This discrepancy can be easily explained by Theorem 2.3 and the fact that only subgradients on $P$ that extend homogeneously to subgradients of $\Phi$ on cone $P$ are PSRs associated with $\Phi$. Thus, we arrive at the following.

**Corollary 2.5.** Consider a (strictly) convex function $\Phi : P \to \mathbb{R}$ that has a subgradient $\Phi^* : P \to \mathcal{L}(P)$ on $P$. Then

$$S(q)(x) = \Phi^*(q)(x) + \Phi(q) - q \cdot \Phi^*(q)$$
is also a (strict) subgradient of $\Phi$ on $\mathcal{P}$, and in addition, the 0-homogeneous extension of $S$ is a (strict) subgradient of the 1-homogeneous extension of $\Phi$ on cone $\mathcal{P}$.

**Proof.** The proof follows immediately from Theorem 2.3 and the fact that $S$ is a PSR due to Theorem 2.4. However, it would be instructive to show the claim independently. The 0-homogeneous extension of $S$ is given by

$$S(q)(x) = \Phi^* \left( \frac{q}{q \cdot 1} \right)(x) + \Phi \left( \frac{q}{q \cdot 1} \right) - \frac{q}{q \cdot 1} \cdot \Phi^* \left( \frac{q}{q \cdot 1} \right).$$  \hspace{1cm} (7)$$

Clearly,

$$q \cdot S(q) = (q \cdot 1) \Phi \left( \frac{q}{q \cdot 1} \right),$$

for any $q \in \text{cone} \mathcal{P}$, as desired. We also have,

$$p \cdot S(q) = p \cdot \Phi^* \left( \frac{q}{q \cdot 1} \right)(x) + (p \cdot 1) \left( \Phi \left( \frac{q}{q \cdot 1} \right) - \frac{q}{q \cdot 1} \cdot \Phi^* \left( \frac{q}{q \cdot 1} \right) \right)$$

$$\leq \left( \left( \frac{p}{p \cdot 1} - \frac{q}{q \cdot 1} \right) \cdot \Phi^* \left( \frac{q}{q \cdot 1} \right) + \Phi \left( \frac{q}{q \cdot 1} \right) \right) (p \cdot 1)$$

$$\leq (p \cdot 1) \Phi \left( \frac{p}{p \cdot 1} \right),$$

for any $p, q \in \text{cone} \mathcal{P}$, as desired.

Another useful consequence from the above characterisations is the following.

**Corollary 2.6.** Let $\Phi : \mathcal{P} \to \mathbb{R}$ be a (strictly) convex function that has a subgradient $\Phi^* : \mathcal{P} \to \mathcal{L}(\mathcal{P})$ on $\mathcal{P}$. Then $\Phi^*$ is a (strictly) PSR associated with $\Phi$ if and only if $q \cdot \Phi^*(q) = \Phi(q)$, for every $q \in \mathcal{P}$.

**Proof.** The proof follows directly from the hypothesis and (6).

Although equivalent, both Theorem 2.3 and Theorem 2.4 have their own relative strengths. As we have said earlier, Theorem 2.3 can be very useful theoretically (Dawid, Lauritzen, and Parry 2012; Ovcharov 2014). On the other hand, Theorem 2.4 is often simpler to use when one needs to find a PSR associated with a given entropy function. We next make a brief illustration of the two theorems. See also Examples 4.3 and 4.4 below.

**Example 2.7.** Let $\mathcal{P}$ denote the set of probability densities in the Lebesgue space $L^2(\mathcal{X}, \mu)$. We consider the quadratic entropy $\Phi(q) = q \cdot q$ on $\mathcal{P}$ and wish to find a PSR associated with $\Phi$. To that end, we extend $\Phi$ as the squared $L^2$-norm on $L^2(\mathcal{X}, \mu) = \text{span} \mathcal{P}$ and make use of the fact that the extended entropy is differentiable. Its functional derivative is given by

$$\Phi^*(q) = 2q.$$
This means that $\Phi^*(q)$ is a subgradient of $\Phi$ on $\mathcal{P}$, however, $\Phi^*$ is not a PSR associated with $\Phi$ as $q \cdot \Phi^*(q) \neq \Phi(q)$. For that reason, we apply Theorem 2.4 to find that

$$S(q) = \Phi^*(q) + \Phi(q) - q \cdot \Phi^*(q) = 2q - q \cdot q$$

is a PSR associated with $\Phi$. This scoring rule is known as the quadratic scoring rule.

On the other hand, let us next consider the spherical entropy on $\mathcal{P}$ defined as $\Phi(q) = (q \cdot q)^{1/2}$, and also find a PSR associated with it. Notice now that $\Phi$ has a natural extension to span $\mathcal{P}$ as the $L^2$-norm, which is a sublinear function. Using the fact that $\Phi$ is a composition of the functions $x \rightarrow x^{1/2}$ and $q \rightarrow q \cdot q$, we find that its functional derivative on span $\mathcal{P}$ is given by

$$\Phi^*(q) = \frac{q}{(q \cdot q)^{1/2}}.$$ 

In the light of either Theorem 2.3 or Corollary 2.6, $\Phi^*$ is a PSR associated with $\Phi$. This scoring rule is known as the spherical scoring rule.

### 3 General convex extensions

Here we explore the case where an entropy function $\Phi : \mathcal{P} \rightarrow \mathbb{R}$ has a convex extension to a set $\mathcal{K}$ such that $\mathcal{P} \subset \mathcal{K} \subset \text{span} \mathcal{P}$. This contains as a special case the extension of $\Phi$ as a sublinear function to cone $\mathcal{P}$ discussed previously. Our immediate goal is to uncover what quantities relate to PSRs in this more general context.

#### 3.1 Functional Bregman divergences

To help explain why we are interested in convex extensions of $\Phi$, we consider another related notion to PSRs – that of divergence. Employing the same notation as above, a divergence on $\mathcal{K}$ is any function $D : \mathcal{K} \times \mathcal{K} \rightarrow \mathbb{R}$ with the following two properties: (i) $D(p, q) \geq 0$, for all $p, q \in \mathcal{K}$, and (ii) $D(p, q) = 0$ implies $p = q$. Divergences are used in statistics to measure distances between probability distributions. They are more general than metrics and need not be symmetric nor satisfy the triangle inequality.

To every scoring rule $S : \mathcal{P} \rightarrow \mathcal{L}(\mathcal{P})$, we associate the function $D : \mathcal{P} \times \mathcal{P} \rightarrow \mathbb{R}$ given by

$$D(p, q) = p \cdot S(p) - p \cdot S(q). \quad (8)$$

It is easy to see that $S$ is proper if and only if $D(p, q) \geq 0$ for all $p, q \in \mathcal{P}$, and $S$ is strictly proper if and only if in addition $D(p, q) = 0$ implies $p = q$. Following Gneiting and Raftery 2007, we call a divergence induced by a PSR a score divergence.

Consider now a PSR $S$ with an associated entropy function $\Phi$ and suppose that $S(q) = \Phi^*(q) - q \cdot \Phi^*(q) + \Phi(q)$ for some subgradient $\Phi^*$ of $\Phi$ on $\mathcal{P}$. Then we may write (8) in the following equivalent way

$$D(p, q) = \Phi(p) - (p - q) \cdot \Phi^*(q) - \Phi(q),$$
for all $p, q \in \mathcal{P}$. Geometrically, for a given $q \in \mathcal{P}$, the condition $D(p, q) \geq 0$, for all $p \in \mathcal{P}$, is equivalent to the condition that the affine functional $p \mapsto (p - q) \cdot \Phi^*(q) - \Phi(q)$ defines a supporting hyperplane to $\Phi$ at $q$. Strict positivity of $D$ corresponds to strict convexity of $\Phi$. Therefore, if we can extend $\Phi$ to a larger domain $\mathcal{K}$ as a strictly convex function and $\Phi^*$ as a subgradient of $\Phi$ on $\mathcal{K}$, that will also extend $D$ as a divergence on $\mathcal{K}$. We refer to such divergences as **functional Bregman divergences**. Functional Bregman divergences have been studied by Frigyik, Srivastava, and Gupta 2008 under different than ours regularity assumptions on $\Phi$.

**Definition 3.1.** Given a strictly convex function $\Phi : \mathcal{K} \to \mathbb{R}$ and a subgradient $\Phi^* : \mathcal{K} \to \mathcal{L}(\mathcal{P})$ of $\Phi$ on $\mathcal{K}$, the functional Bregman divergence on $\mathcal{K}$ associated with the pair $(\Phi, \Phi^*)$ is the function $D_{(\Phi, \Phi^*)} : \mathcal{K} \times \mathcal{K} \to \mathbb{R}$ given by

$$D_{(\Phi, \Phi^*)}(p, q) = \Phi(p) - (p - q) \cdot \Phi^*(q) - \Phi(q),$$

for all $p, q \in \mathcal{K}$.

So, for sets of probability densities $\mathcal{P}$, score divergences and functional Bregman divergences coincide, however, the latter yield natural generalisations of the former on extended domains. The relation (9) motivates defining the function $s : \mathcal{K} \times \mathcal{K} \to \mathbb{R}$, $s(p, q) = (p - q) \cdot \Phi^*(q) + \Phi(q),$ for all $p, q \in \mathcal{K}$. This allows us to write (9) as

$$D_{(\Phi, \Phi^*)}(p, q) = s(p, p) - s(p, q)$$

for all $p, q \in \mathcal{K}$. Furthermore, when restricting $s$ to $\mathcal{P} \times \mathcal{P}$, we may put it in the form

$$s(p, q) = p \cdot S(q),$$

where $S(q) = \Phi^*(q) - q \cdot \Phi^*(q) + \Phi(q)$ is a PSR. Thus, we may view the function $s$ as an extension of the expected score of $S$ to $\mathcal{K} \times \mathcal{K}$ that preserves the associated functional Bregman divergence. We call $s$ an **extended score function** associated with $S$ for a general convex set $\mathcal{K}$, while for $\mathcal{P}$, we refer to $s$ as the **score function** associated with $S$.

To formalise the concept of an extended score function, we need to introduce the following class of functionals. By $\mathcal{A}(\mathcal{P})$ we denote the vector space of affine functionals $A$ on span $\mathcal{P}$ of the form $A(p) = p \cdot f + \phi(p)$, where $f \in \mathcal{L}(\mathcal{P})$ and $\phi : \text{span} \mathcal{P} \to \mathbb{R}$.

**Definition 3.2.** We say that $s : \mathcal{K} \times \mathcal{K} \to \mathbb{R}$ is an **extended score function** if for every $q \in \mathcal{K}$, $s(\cdot, q) \in \mathcal{A}(\mathcal{P})$. Furthermore, $s$ is said to be (strictly) proper if $s$ (strictly) satisfies

$$s(p, q) \leq s(p, p)$$

for all $p, q \in \mathcal{K}$. 
The above definition is analogous to the definition of an affine score by Frongillo and Kash 2014 introduced by them in a slightly different context. Notice that if $s$ is proper, the function $p \to s(p, p)$ is convex as maximum of affine functions. We next characterise the extended score functions that are (strictly) proper.

**Theorem 3.3.** An extended score function $s : \mathcal{K} \times \mathcal{K} \to \mathbb{R}$ is (strictly) proper if and only if there is a (strictly) convex function $\Phi : \mathcal{K} \to \mathbb{R}$ and a subgradient $\Phi^* : \mathcal{K} \to \mathcal{L}(\mathcal{P})$ of $\Phi$ on $\mathcal{K}$ such that

$$s(p, q) = (p - q) \cdot \Phi^*(q) + \Phi(q)$$

for all $p, q \in \mathcal{K}$.

The result is similar to Frongillo and Kash 2014, Theorem 1, however, due to differences in notation and regularity assumptions, we sketch a proof.

**Proof.** Suppose that $s$ is an extended score function that is proper. Then the function $\Phi(p) = s(p, p)$, for $p \in \mathcal{K}$, is convex. By hypothesis, $s$ is affine in its first argument. Hence, for each $q \in \mathcal{K}$, we may write $s$ in the form $s(p, q) = (p - q) \cdot q^* + \Phi(q)$, where $q^* \in \mathcal{L}(\mathcal{P})$. From the fact that $s$ is proper it follows that $\Phi(p) \geq (p - q) \cdot q^* + \Phi(q)$, for all $p \in \mathcal{K}$, and hence, $q^*$ is a subgradient of $\Phi$ at $q$. Setting $\Phi^*(q) = q^*$, the necessity part of the claim follows. Conversely, if $\Phi$ is convex and $\Phi^*$ is a subgradient of $\Phi$ on $\mathcal{K}$, then, for each $q \in \mathcal{K}$, the graph of the affine functional $p \to s(p, q)$ defined in (10) is a supporting hyperplane to $\Phi$ at $q$. This is equivalent to $s$ being proper.

For the case of strict propriety, let us recall that a convex function $\Phi$ is strictly convex if, for every two different points $p, q \in \mathcal{P}$, no subgradient $q^*$ of $\Phi$ at $q$ is a subgradient of $\Phi$ at $p$. For $s$ in the form (10), this is equivalent to $s$ being strictly proper.

Theorem 3.3 generalises Theorem 2.3 in that it describes general convex extensions of entropy functions and relates to them proper score functions – the natural extension of the notion of a PSR in the present context. In the following corollary, we describe the important special case where the score functions are linear in their first argument and may be defined as expectations of subgradients of entropy functions outside cone $\mathcal{P}$. To that end, let $\mathcal{C}$ denote a convex cone such that $\mathcal{P} \subset \mathcal{C} \subset \text{span}(\mathcal{P})$.

**Corollary 3.4.** Let $s : \mathcal{C} \times \mathcal{C} \to \mathbb{R}$ be an extended score function that is proper and let $\Phi : \mathcal{C} \to \mathbb{R}$, $\Phi(p) = s(p, p)$, be the associated extended entropy. Then, $s$ is linear in its first argument if and only if $\Phi$ is 1-homogeneous on $\mathcal{C}$.

To summarise, this subsection describes a basic convex construction that incorporates PSRs and Bregman divergences in a natural way. For simplicity, let us call a pair $(\Phi, \Phi^*)$ admissible whenever $\Phi : \mathcal{K} \to \mathbb{R}$ is a convex function and $\Phi^* : \mathcal{K} \to \mathcal{L}(\mathcal{P})$ is a subgradient of $\Phi$ on $\mathcal{K}$. So, given an admissible pair $(\Phi, \Phi^*)$, we have shown above how one may define a functional Bregman divergence, an
Extensions of PSRs

11

extended score function, and a PSR associated to \((\Phi, \Phi^*)\), and we have explained geometrically the relation between these quantities. The converse is also true: One may start with a PSR \(S\) and then define an associated entropy function \(\Phi\), and through the convex extensions of \(\Phi\), one may define associated admissible pairs \((\Phi, \Phi^*)\).

3.2 Properties of functional Bregman divergences

Here we describe some of the basic properties of functional Bregman divergences under natural regularity conditions for PSRs. The following properties extend similar results from more specialised contexts and have almost identical proofs, which we omit (Banerjee et al. 2005; Frigyik, Srivastava, and Gupta 2008).

1. \(D_{(\Phi, \Phi^*)}(p, q)\) is convex in its first argument.

2. The set of admissible pairs is a convex cone: given two admissible pairs \((\Phi_1, \Phi_1^*)\) and \((\Phi_2, \Phi_2^*)\), and two constants \(c_1, c_2 \geq 0\), the pair \((c_1\Phi_1 + c_2\Phi_2, c_1\Phi_1^* + c_2\Phi_2^*)\) is also admissible. Moreover,
\[
D_{(c_1\Phi_1 + c_2\Phi_2, c_1\Phi_1^* + c_2\Phi_2^*)} = c_1D_{(\Phi_1, \Phi_1^*)} + c_2D_{(\Phi_2, \Phi_2^*)}.
\]

3. \(D_{(\Phi_1, \Phi_1^*)}\) and \(D_{(\Phi_2, \Phi_2^*)}\) are identical whenever \(\Phi_1 - \Phi_2\) is an affine function in \(A(P)\).

4. Linear separation. Given \(a, b \in \text{span} \mathcal{P}\), the solutions \(f \in \text{span} \mathcal{P}\) to the equation
\[
D_{(\Phi, \Phi^*)}(a, f) = D_{(\Phi, \Phi^*)}(b, f)
\]
form a hyperplane in \(\text{span} \mathcal{P}\).

5. Generalised Pythagorean theorem. For any three points \(p_1, p_2, p_3 \in \text{span} \mathcal{P}\), the following property holds:
\[
D_{(\Phi, \Phi^*)}(p_1, p_3) = D_{(\Phi, \Phi^*)}(p_1, p_2) + D_{(\Phi, \Phi^*)}(p_2, p_3) - (p_1 - p_2) \cdot (\Phi^*(p_3) - \Phi^*(p_2)).
\]

We next characterise functional Bregman divergences. The result extends a similar claim in Banerjee et al. 2005, Appendix A from the Euclidean setting. We thus also address a question of Gneiting and Raftery 2007 calling for characterisation of score divergences.

**Theorem 3.5.** Let \(D : \mathcal{K} \times \mathcal{K} \to \mathbb{R}\) be a divergence on \(\mathcal{K}\). Then \(D\) is a functional Bregman divergence on \(\mathcal{K}\) if and only if for any \(a \in \mathcal{K}\) the function \(\Phi(p) = D(p, a)\) is strictly convex and \(\Phi\) has a subgradient \(\Phi^* : \mathcal{K} \to \mathcal{L}(\mathcal{P})\) such that
\[
D(p, q) = D_{(\Phi, \Phi^*)}(p, q)
\]
for all \(p, q \in \mathcal{K}\). Here \(D_{(\Phi, \Phi^*)}\) is the functional Bregman divergence associated with the pair \((\Phi, \Phi^*)\).
Proof. Suppose that $D$ is a functional Bregman divergence associated with some admissible pair $(\Phi_1, \Phi_1^*)$. Then,

$$
\Phi(q) = \Phi_1(p) - p \cdot \Phi_1^*(a) + a \cdot \Phi_1^*(a) - \Phi_1(a).
$$

Since $\Phi(q)$ and $\Phi_1(p)$ only differ by an element in $\mathcal{A}(\mathcal{P})$, the necessity part of the claim follows. The sufficiency part is trivial.

A divergence function $D$ on $\mathcal{K}$ is said to be symmetric whenever $D(p, q) = D(q, p)$ for all $p, q \in \mathcal{K}$. The last point we consider in this subsection is to characterise the symmetric functional Bregman divergences associated with a broad class of entropy functions $\Phi$. To that end, let us first recall certain common families of functional Bregman divergences.

Let $f : [0, \infty) \to \mathbb{R}$ be a convex and differentiable function. The function $D_f : [0, \infty) \times [0, \infty) \to \mathbb{R}$ given by

$$
D_f(x, y) = f(x) - f'(y)(x - y) - f(y)
$$

is the scalar Bregman divergence on $[0, \infty)$ associated with $f$. We now use $D_f$ to define a functional Bregman divergence on $\mathcal{K}$ as follows. Let $\nu$ be a measure on $\mathcal{X}$ that is absolutely continuous with respect to $\mu$ and consider the function $D : \mathcal{K} \times \mathcal{K} \to \mathbb{R}$ given by

$$
D(p, q) = \int_{\mathcal{X}} D_f(p(x), q(x))d\nu(x)
$$

for any $p, q \in \mathcal{K}$. The divergences $D$ of this form are commonly known as separable Bregman divergences. For example, choosing $f(x) = x^\gamma$, for $1 < \gamma < \infty$, we obtain the important family of separable Bregman divergences associated with the power entropy,

$$
\Phi(p) = \int_{\mathcal{X}} p^\gamma(x)d\nu(x).
$$

We now include the separable Bregman divergences into a larger family of functional Bregman divergences. To that end, let $\Phi : \mathcal{K} \to \mathbb{R}$ be a strictly convex function of the form

$$
\Phi(p) = \phi \left( \int_{\mathcal{X}} f(p(x))d\nu(x) \right),
$$

where $f$ and $\nu$ are as above, while $\phi : \mathbb{R} \to \mathbb{R}$ is an increasing function. The subfamily of separable Bregman divergences correspond to the identity function $\phi(x) = x$. An important subfamily of (non-separable) divergences is obtained by choosing $f(x) = x^\gamma$ and $\phi(x) = x^{1/\gamma}$, for $1 < \gamma < \infty$. These divergences are associated with the pseudospherical entropy,

$$
\Phi(p) = \left( \int_{\mathcal{X}} p^\gamma(x)d\nu(x) \right)^{1/\gamma}.
$$
Extensions of PSRs

This example shows that $\Phi$ may be convex without $\phi$ being convex. However, not all $\Phi$ in the form (11) will be convex. A simple condition guaranteeing that is to assume that $\phi$ is convex and increasing.

We are now ready to characterise the symmetric Bregman divergences associated with entropies of the form (11). The proof extends an analogous result in the scalar case (Bauschke and Borwein 2001; Iusem 1992). To that end, let us suppose that span $\mathcal{P}$ may be identified with a Fréchet space $\mathcal{N}$, and let $\mathcal{K}$ be an open convex set in $\mathcal{N}$ containing $\mathcal{P}$. We denote by $\mathcal{N}^*$ the topological dual space of $\mathcal{N}$, which we assume to be identifiable with a subspace of $\mathcal{L}(\mathcal{P})$.

Theorem 3.6. With the above notations, let $\Phi : \mathcal{K} \to \mathbb{R}$ be a strictly convex function of the form (11). Suppose also that $\phi$ and $f$ are twice differentiable and $\Phi$ is twice Fréchet differentiable. If the associated functional Bregman divergence is symmetric, then $\Phi$ has the form

$$\Phi(q) = q \cdot \rho q$$

or

$$\Phi(q) = (q \cdot \rho)^2,$$

up to affine terms of the form $\alpha q \cdot \rho + \beta$, where $\rho : \mathcal{N} \to \mathbb{R}$ is the density of the measure $\nu$ with respect to $\mu$, and $\alpha, \beta \in \mathbb{R}$ are constants.

Proof. We give a sketch of the proof. Let $\Phi' : \mathcal{N} \to \mathbb{R}$ and $\Phi'' : \mathcal{N} \times \mathcal{N} \to \mathbb{R}$ denote the first and second Fréchet derivatives of $\Phi$ on $\mathcal{K}$. A computation shows

$$\xi \cdot \Phi'(p) = \phi'(\int_{\mathcal{N}} f(p(x))d\nu(x)) \int_{\mathcal{N}} f'(p(x))\xi(x)d\nu(x),$$

$$(\xi, \eta) \cdot \Phi''(p) = \phi''\left(\int_{\mathcal{N}} f(p(x))d\nu(x)\right) \int_{\mathcal{N}} f''(p(x))\xi(x)\eta(x)d\nu(x) + \phi''\left(\int_{\mathcal{N}} f'(p(x))\xi(x)d\nu(x)\right) \int_{\mathcal{N}} f'(p(x))\eta(x)d\nu(x).$$

We remark that “$\cdot$” denotes both the duality pairing with respect to $\mathcal{N}$ and $\mathcal{N}^*$, and with respect to span $\mathcal{P}$ and $\mathcal{L}(\mathcal{P})$. This is well-justified if we have that $f'(p)\rho$ and $f''(p)\rho\xi$ are in $\mathcal{L}(\mathcal{P})$ for all $p \in \mathcal{K}$ and all $\xi \in \text{span} \mathcal{P}$.

Symmetry of the Bregman divergence associated with $\Phi$ means that we have the identity

$$2\Phi(p) - (p - q) \cdot \Phi'(q) = 2\Phi(q) - (q - p) \cdot \Phi'(p)$$

Let $p_t$ denote $p + tr$, for $t \in [0, 1]$, $r \in \mathcal{N}$. Replace $p$ with $p_t$ above and differentiate with respect to $t$ at $t = 0$ to find

$$2r \cdot \Phi'(p) - r \cdot \Phi'(q) = r \cdot \Phi'(p) - ((q - p), r) \cdot \Phi''(p).$$

Since $r \in \mathcal{N}$ is arbitrary, we have

$$\Phi'(q) = (q, \cdot) \cdot \Phi''(p) + \Phi'(p) - (p, \cdot) \cdot \Phi''(p).$$
Fix $p$ and consider that $q$ is the only variable above. Using the explicit form of $\Phi''(p)$, we get that
\[ \Phi'(q) = 2\alpha a q + 2\beta (q \cdot b) b + c, \]
where $\alpha, \beta \in \mathbb{R}$, and $a, b, c : \mathcal{X} \to \mathbb{R}$. In view of the fundamental theorem of calculus for Fréchet spaces (Hamilton 1982, Theorem 3.2.2),
\[ \Phi(q) = \alpha q \cdot a q + \beta (q \cdot b)^2 + q \cdot c + \gamma, \]
where $\gamma$ is a constant of integration. (The latter claim may be verified directly by differentiation.) Since $\Phi$ must be in the form (11), the claim follows.

In view of the Theorem, the only symmetric functional Bregman divergences on $\mathcal{K}$ induced by convex functions $\Phi$ in the form (11) are the following:
\[ D_1(p, q) = \int_{\mathcal{X}} (p(x) - q(x))^2 d\nu(x), \]
\[ D_2(p, q) = \left( \int_{\mathcal{X}} (p(x) - q(x)) d\nu(x) \right)^2. \]
Notice that by the Cauchy-Schwartz inequality,
\[ D_2(p, q) \leq \int_{\mathcal{X}} 1 d\nu(x) D_1(p, q), \]
hence the quadratic divergence $D_1$ is essentially the only symmetric divergence associated with entropy functions having the form (11).

4 Subdifferentials of entropy functions

Our goal here is to describe all PSRs that are associated with a given entropy function $\Phi$. We also show in examples how one may find explicitly a PSR that is associated with a given $\Phi$. As it is well-known, subdifferentials and directional derivatives are closely related. We begin by introducing some notions and properties related to directional derivatives.

A direction in a vector space is the equivalence class of all positively collinear vectors to a given nonzero vector and is represented by any member of this class. For a convex set $\mathcal{K}$, $\mathcal{P} \subset \mathcal{K} \subset \text{span} \mathcal{P}$, and a point $q \in \mathcal{K}$, the set of directions at $q$ that point towards $\mathcal{K}$ is cone($\mathcal{K} - q$) (Ovcharov 2014, Lemma 2.1).

**Definition 4.1.** The right directional derivative of $\Phi : \mathcal{K} \to \mathbb{R}$ at $q \in \mathcal{K}$ along the vector $p \in \text{cone}(\mathcal{K} - q)$ is defined as the limit
\[ \Phi'_+(p, q) = \lim_{t \to 0^+} \frac{\Phi(q + tp) - \Phi(q)}{t}, \]
whenever it exists.
When \( \Phi \) is convex, the above limit is equivalent to
\[
\Phi_+(p, q) = \inf_{t > 0} \frac{\Phi(q + tp) - \Phi(q)}{t},
\]
and hence it always exists but may be equal to \(-\infty\). Moreover, the limit is finite whenever \( q \) lies in the (relative) interior of a line segment in \( K \) with direction \( p \) (Rockafellar 1972; Aragón Artacho et al. 2014). The directions at \( q \) with that property are clearly given by the set \( \mathcal{O}(q) = \text{cone}(K - q) \cap \text{cone}(K - q) \), which is a vector subspace of span \( P \). Replacing \( t \to 0^+ \) in the limit \((12)\) with \( t \to 0^- \), we obtain the left directional derivative, denoted \( \Phi_-(p, q) \), for \( p \in \mathcal{O}(q) \). It is easy to see that \( \Phi_-(p, q) = -\Phi_+(p, q) \), for all \( p \in \mathcal{O}(q) \). Whenever, for some \( p \in \mathcal{O}(q) \), \( \Phi_-(p, q) = \Phi_+(p, q) \), we say that \( \Phi \) has a two-sided (or bilateral) directional derivative along \( p \) at \( q \), denoted \( p \cdot \Phi'(q) \). This notation may be justified by the fact that the two-sided directional derivative is a linear functional defined on a vector subspace of \( \mathcal{O}(q) \). If, for some \( q \in K \), \( \mathcal{O}(q) = \text{span} P \), we say that \( q \) is an algebraically interior point of \( K \). If, on the other hand, \( \mathcal{O}(q) \neq \text{span} P \), \( q \) is a boundary point of \( K \). Whenever \( q \) is a boundary point of \( K \) but the space \( \mathcal{O}(q) \) is sufficiently large in a certain sense, various weaker notions of interior point have been devised under the common term quasi-interior point (Fullerton and Braunschweiger 1963; Borwein and Lewis 1992).

**Theorem 4.2.** Let \( \Phi : K \to \mathbb{R} \) be a convex function. Then \( \Phi \) has a \( P \)-integrable subgradient at a point \( q \in K \) if and only if there is \( q^* \in L(P) \) such that
\[
p \cdot q^* \leq \Phi_+(p, q)
\]
for all \( p \in K \), with equality for \( p = q \).

**Proof.** The proof is a minor variant of Ovcharov 2014, Theorem 3.1. \( \square \)

The necessary and sufficient condition for subgradient in the theorem above improves greatly the analogous condition in the definition of subgradient. It is well-known that directional derivatives are sublinear in \( p \) and hence the above inequality relates a linear map \( q^* \) to a sublinear map \( \Phi_+(\cdot, q) \). Often in practice \( \Phi_+(\cdot, q) \) turns out to be linear, and then \( \Phi_+(\cdot, q) \) is a subgradient provided it meets our regularity conditions.

**Example 4.3.** Let us consider the sublinear extension of the Shannon entropy function for probability densities on \( \mathbb{R}^d \),
\[
\Phi(p) = \int_{\mathbb{R}^d} p(x) \ln \frac{p(x)}{p \cdot 1} dx.
\]
A suitable set of denormalised densities for which the above integral is convergent is given by
\[
\text{cone} \mathcal{P} = \left\{ p \in C(\mathbb{R}^d) \mid p(x) > 0, \exists C_1, C_2 > 0 : \frac{C_1}{(1 + |x|)^d} \leq p(x) \leq \frac{C_2}{(1 + |x|)^{d+1}} \right\},
\]
where \( C(\mathbb{R}^d) \) denotes the space of continuous functions on \( \mathbb{R}^d \) and \( \delta \geq d + 1 \) is arbitrary. Since \( \text{span} \ P = \text{cone} \ P - \text{cone} \ P \), it follows that

\[
\text{span} \ P = \left\{ f \in C(\mathbb{R}^d) \mid \exists C_1, C_2 > 0 : \frac{C_1}{(1 + |x|)^{\delta}} \leq |f(x)| \leq \frac{C_2}{(1 + |x|)^{d+1}} \right\}.
\]

It is easy to show that \( L(P) \) may be identified with a subspace of \( L^1_{\text{loc}}(\mathbb{R}^d) \), the locally Lebesgue integrable functions on \( \mathbb{R}^d \) (Ovcharov 2014, Section 4.1). Let us find, for each \( q \in \text{cone} \ P \), the space \( O(q) = \text{cone}(P - q) \cap -\text{cone}(P - q) \). It follows straight from definitions that \( f \in \text{span} \ P \) is in \( O(q) \) if and only if there are constants \( \alpha_{\pm} > 0 \) such that \( q \pm \alpha_{\pm} f \in \text{cone} \ P \). The functions satisfying the latter condition are precisely those whose rate of decay at infinity is either equal or faster than that of \( q \). Therefore, the subspace \( O(q) \) varies with \( q \in \text{span} \ P \) and \( O(q) \subsetneq \text{span} \ P \). Notice as well that \( \text{span} \ P \) and \( O(q) \), for each \( q \in \text{cone} \ P \), are dense in \( L^1(\mathbb{R}^d) \).

A simple computation (Ovcharov 2014, Section 4.1) shows that

\[
\Phi_{\pm}'(p, q) = \int_{\mathbb{R}^d} p(x) \ln \frac{q(x)}{q \cdot 1} dx.
\]

In view of Theorem 4.2,

\[
S(q)(x) = \ln \frac{q(x)}{q \cdot 1}
\]

is a subgradient of \( \Phi \) at \( q \) on \( \text{cone} \ P \). The inspection that \( S(q) \) is a \( \mathcal{P} \)-integrable function is straightforward. We recognise in \( S(q) \) the 0-homogeneous extension of the logarithmic scoring rule. Since \( \Phi_{\pm}'(p, q) \) is linear in \( p \) on the space \( \mathcal{O}(q) \), it follows that \( \Phi \) has a two-sided directional derivative \( p \cdot \Phi_{\pm}'(q) = \Phi_{\pm}'(p, q) \) for all \( p \in \mathcal{O}(q) \). Therefore, the logarithmic scoring rule may be identified with \( \Phi_{\pm}'(q) \) on the subspaces \( \mathcal{O}(q) \). In a certain sense the directions in the complement of \( \mathcal{O}(q) \) with respect to \( \text{span} \ P \) form a negligible set. More specifically, any two linear functionals in \( L(P) \) that agree on \( \mathcal{O}(q) \) are identical. In addition, \( \Phi \) has a unique supporting hyperplane at each point in \( \text{cone} \ P \) with subgradient in \( L(P) \) (Ovcharov 2014). Note that this holds without \( \Phi \) being differentiable. The example gives a precise sense in which we may regard \( S \) and \( \Phi \) to be uniquely associated with each other.

The same techniques as in the above example are especially useful in the context of proper local scoring rules of higher orders as their entropies are in general non-differentiable. See Parry, Dawid, and Lauritzen 2012; Ehm and Gneiting 2012; Forbes and Lauritzen 2014 for some results about the latter family of scoring rules. For comparison purposes, let us next consider a non-homogeneous extension of the Shannon entropy function and find an associated PSR.

**Example 4.4.** With the same notation as in the previous example, consider

\[
\Phi(p) = \int_{\mathbb{R}^d} p(x) \ln p(x) dx
\]

(14)
Extensions of PSRs

for \( p \in \text{cone} \mathcal{P} \). Let us compute the right directional derivative of \( \Phi \). For \( q \in \text{cone} \mathcal{P} \) and \( p \in \text{cone}(\text{cone} \mathcal{P} - q) \), we set \( q_t = q + tp \). We have

\[
\lim_{t \to 0^+} \frac{\Phi(q + tp) - \Phi(q)}{t} = \frac{d}{dt} \bigg|_{t=0} (q_t \cdot \ln q_t) = p \cdot \ln q + \frac{p}{q} = p \cdot \ln q + p \cdot 1.
\]

Thus we obtain that \( \Phi^*(q) = \ln q + 1 \) is a subgradient of \( \Phi \). Note that although \( \Phi^* \) is a PSR, as it differs by a PSR by a constant, it is not a PSR that is associated with \( \Phi \), as \( q \cdot \Phi^*(q) \neq \Phi(q) \). Applying Theorem 2.4, we obtain that

\[ S(q) = \Phi^*(q) - q \cdot \Phi^*(q) + \Phi(q) = \ln q \]

is a PSR associated with \( \Phi \).

We now proceed to the case where the extended entropy is a differentiable function. We suppose that \( \text{span} \mathcal{P} \) may be identified with a normed space \((\mathcal{N}, \|\cdot\|)\) and denote by \( \mathcal{N}^* \) the topological dual space of \( \mathcal{N} \). We further assume that \( \mathcal{N}^* \) may be identified with a subspace of \( L(\mathcal{P}) \).

**Definition 4.5.** Let \( K \) be an open convex set in \( \mathcal{N} \). A function \( \Phi : K \to \mathbb{R} \) is **Gâteaux differentiable** at a point \( q \in K \) if there is \( q^* \in \mathcal{N}^* \) such that for every \( p \in \mathcal{N} \), the limit

\[ p \cdot q^* = \lim_{t \to 0} \frac{\Phi(q + tp) - \Phi(q)}{t} \]

exists. The functional \( q^* \) is called the **Gâteaux derivative** of \( \Phi \) at \( q \).

The Gâteaux derivative is necessarily unique from definition. We say that \( \Phi \) is differentiable on \( K \) if \( \Phi \) is differentiable at every point in \( K \). The following is a standard result from convex analysis (Aragón Artacho et al. 2014; Borwein and Vanderwerff 2010; Zalinescu 2002).

**Theorem 4.6.** Let \( K \) be an open convex set in \( \mathcal{N} \) and \( \Phi : K \to \mathbb{R} \) be a convex and continuous function. Then, \( \Phi \) is Gâteaux differentiable on \( K \) if and only if \( \Phi \) admits a unique subgradient \( \Phi^* : K \to \mathcal{N}^* \) on \( K \). In this case \( \Phi^* \) is the Gâteaux derivative of \( \Phi \) on \( K \).

In the light of the theorem, every convex differentiable function \( \Phi : K \to \mathbb{R} \) with gradient \( \Phi^* : \mathcal{N} \to \mathcal{N}^* \), where \( \mathcal{P} \subset K, \mathcal{N}^* \subset L(\mathcal{P}) \), defines a unique collection of supporting hyperplanes to its graph. Hence, there is a unique extended score function \( s \) on \( K \) that is associated with \( \Phi \). Restricted to \( \mathcal{P} \), \( s \) is the score function of the PSR \( S : \mathcal{P} \to L(\mathcal{P}) \), given by \( S(q) = \Phi^*(q) - q \cdot \Phi^*(q) + \Phi(q) \). We cannot claim that \( S \) and \( \Phi \) are uniquely associated with each other, unless we identify \( S \) with \( s \), which is often done in practice.

**Example 4.7.** Let \( \mathcal{P} \) be the set of all probability densities in \( \mathcal{N} = L^\gamma(X, \mu) \), for \( 1 < \gamma < \infty \), and consider the power entropy function,

\[ \Phi_\gamma(p) = \int_X p^\gamma(x) d\mu(x), \]
for $p \in L^\gamma(X, \mu)$. We have that span $P = N$ and the topological dual space of $L^\gamma(X, \mu)$ is $N^* = L^{\gamma/(\gamma-1)}(X, \mu)$. Clearly, $N^*$ may be identified with a subspace of $\mathcal{L}(P)$.

We proceed to compute the Gâteaux derivative of $\Phi_\gamma$ on $L^\gamma(X, \mu)$. We have

$$\lim_{t \to 0} \frac{\Phi_\gamma(q + tp) - \Phi_\gamma(q)}{t} = \left. \frac{d}{dt} \right|_{t=0} (q_t)^\gamma \cdot 1$$

$$= p \cdot q^{\gamma-1}.$$  

Since $\Phi_\gamma^*(q) = \gamma q^{\gamma-1} \in N^*$, $\Phi_\gamma^*$ is the Gâteaux derivative of $\Phi_\gamma$. Thus, the Bregman divergence on $L^\gamma(X, \mu)$ associated with $\Phi_\gamma$ is:

$$D_\gamma(p, q) = p \cdot q^{\gamma-1} - (p - q) \cdot \gamma q^{\gamma-1} - q \cdot q^{\gamma-1}.$$  

The associated extended score function is:

$$s_\gamma(p, q) = (p - q) \cdot q^{\gamma-1} + q \cdot q^{\gamma-1}.$$  

Then, the restriction of $s_\gamma$ to $P$ yields the PSR

$$S_\gamma(q) = \gamma q^{\gamma-1} - (\gamma - 1)q \cdot q^{\gamma-1},$$  

the power scoring rule with exponent $\gamma$. Since it seems natural to identify $S_\gamma$ with $s_\gamma$, in this sense we may regard that $S_\gamma$ and $\Phi_\gamma$ are uniquely associated with each other.

References

Abernethy, J. D. and R. M. Frongillo (2012). “A characterization of scoring rules for linear properties”. In: *JMLR: Workshop and Conference Proceedings* 23, pp. 27.1–27.13.

Aragón Artacho, F. J., J. M. Borwein, V. Martín-Márquez, and L. Yao (2014). “Applications of convex analysis within mathematics”. In: *Mathematical Programming* 148, pp. 49–88.

Banerjee, A., S. Merugu, I. S. Dhillon, and J. Ghosh (2005). “Clustering with Bregman divergences”. In: *Journal of Machine Learning Research* 6, pp. 1705–1749.

Bauschke, H. H. and J. M. Borwein (2001). “Joint and separate distance of the Bregman distance”. In: *Studies in Computational Mathematics* 8, pp. 23–36.

Borwein, J. and J. Vanderwerff (2010). “Convex Functions: Constructions, Characterizations and Counterexamples”. In: Encyclopedia of Mathematics and its Applications 109. Cambridge: Cambridge University Press.

Borwein, J. M. and A. S. Lewis (1992). “Partially finite convex programming, Part I: quasi relative interiors and duality theory”. In: *Mathematical Programming* 57.1–3, pp. 15–48.

Brier, G. W. (1950). “Verification of forecasts expressed in terms of probability”. In: *Monthly Weather Review* 78.1, pp. 1–3.
REFERENCES

Bröcker, J. (2009). “Reliability, sufficiency, and the decomposition of proper scores”. In: Quarterly Journal of the Royal Meteorological Society 135, pp. 1512–1519.

Dawid, A. P., S. Lauritzen, and M. Parry (2012). “Proper local scoring rules on discrete sample spaces”. In: The Annals of Statistics 40.1, pp. 593–608.

Dawid, A. P. and M. Musio (2012). “Estimation of spatial processes using local scoring rules”. In: ASA Advances in Statistical Analysis 96. Spatial special issue, pp. 1–7.

— (2014). “Theory and applications of proper scoring rules”. In: Metron 72, pp. 169–183.

Ehm, W. and T. Gneiting (2012). “Local proper scoring rules of order two”. In: The Annals of Statistics 40.1, pp. 609–637.

Forbes, P. G. M. and S. Lauritzen (2014). “Linear estimating equations for exponential families with application to Gaussian linear concentration models”. In: Linear Algebra and its Applications. In press.

Frigyik, B. A., S. Srivastava, and M. R. Gupta (2008). “Functional Bregman divergence and Bayesian estimation of distributions”. In: IEEE Transactions on Information Theory 54, pp. 5130–5139.

Frongillo, R. M. and I. Kash (2014). “Properties in elicitation”. In: available in arxiv.org.

Fullerton, R. E. and C. C. Braunschweiger (1963). “Quasi-interior points of cones”. In: Technical Report 2, University of Delaware, Newark, Delaware.

Gneiting, T. and M. Katzfuss (2014). “Probabilistic forecasting”. In: Annual Review of Statistics and Its Application 1, pp. 125–151.

Gneiting, T. and A. E. Raftery (2007). “Strictly proper scoring rules, prediction, and estimation”. In: Journal of the American Statistical Association 102, pp. 359–376.

Good, I. J. (1952). “Rational decisions”. In: Journal of the Royal Statistical Society, Ser. B 14.1, pp. 107–114.

Hamilton, R. S. (1982). “The inverse function theorem of Nash and Moser”. In: Bulletin of the American Mathematical Society 7.1, pp. 65–222.

Hendrickson, A. and R. Buehler (1971). “Proper scores for probability forecasters”. In: The Annals of Mathematical Statistics 42.6, pp. 1916–1921.

Hyvärinen, A. (2005). “Estimation of non-normalized statistical models by score matching”. In: Journal of Machine Learning Research 6, pp. 695–709.

— (2007). “Some extensions of score matching”. In: Computational Statistics & Data Analysis 51, pp. 2499–2512.

Iusem, A. N. (1992). “A short convergence proof of the EM algorithm for a specific Poisson model”. In: Brazilian Journal of Probability and Statistics 6.1, pp. 57–68.

Kanamori, T. and H. Fujisawa (2014). “Robust estimation under heavy contamination using enlarged models”. In: Biometrika. In press.

Lambert, N. (2013). “Elicitation and evaluation of statistical forecasts”. In: Working Paper.

McCarthy, J. (1956). “Measures of the value of information”. In: Proceedings of the National Academy of Sciences 42, pp. 654–655.
Ovcharov, E. Y. (2014). “Existence and uniqueness of proper scoring rules”. In: Journal of Machine Learning. In press.
Parry, M., A. P. Dawid, and S. Lauritzen (2012). “Proper local scoring rules”. In: The Annals of Statistics 40.1, pp. 561–592.
Rockafellar, R. T. (1972). “Convex Analysis”. In: Second edition. Princeton Mathematical Series. Princeton University Press.
Savage, L. J. (1971). “Elicitation of personal probabilities and expectations”. In: Journal of the American Statistical Association 66, pp. 783–801.
Steinwart, I., C. Pasin, R. C. Williamson, and S. Zhang (2014). “Elicitation and Identification of Properties”. In: Journal of Machine Learning: Workshop and Conference Proceedings 35, pp. 482–526.
Williamson, R. C. (2014). “The geometry of losses”. In: JMLR: Workshop and Conference Proceedings 35, pp. 1078–1108.
Zalinescu, C. (2002). Convex Analysis in General Vector Spaces. World Scientific.
Ziegel, J. F. (2014). “Coherence and elicitability”. In: Mathematical Finance. Preprint.