Existence and Uniqueness of Proper Scoring Rules

Evgeni Y. Ovcharov   TRULR6@YAHOO.COM
Heidelberg Institute for Theoretical Studies
Schloss-Wolfsbrunnenweg 35, D-69118 Heidelberg
Germany

Editor: Zhihua Zhang

Abstract

To discuss the existence and uniqueness of proper scoring rules one needs to extend the associated entropy functions as sublinear functions to the conic hull of the prediction set. In some natural function spaces, such as the Lebesgue L^p-spaces over $\mathbb{R}^d$, the positive cones have empty interior. Entropy functions defined on such cones have directional derivatives only, which typically exist on large subspaces and behave similarly to gradients. Certain entropies may be further extended continuously to open cones in normed spaces containing signed densities. The extended entropies are Gâteaux differentiable except on a negligible set and have everywhere continuous subgradients due to the supporting hyperplane theorem. We introduce the necessary framework from analysis and algebra that allows us to give an affirmative answer to the titular question of the paper. As a result of this, we give a formal sense in which entropy functions have uniquely associated proper scoring rules. We illustrate our framework by studying the derivatives and subgradients of the following three prototypical entropies: Shannon entropy, Hyvärinen entropy, and quadratic entropy.

Keywords: proper scoring rules, entropy, characterisation, existence, uniqueness, quasi-interior, directional derivative, Gâteaux derivative, subgradient, sublinear, convex analysis

1. Introduction

Proper scoring rules have attracted a lot of interest in recent years in disparate fields such as statistics, decision theory, machine learning, game theory, finance, meteorology, etc. They provide practical measures for assessing the accuracy and precision of probabilistic forecasts. In this paper, we build a general measure-theoretic framework for proper scoring rules that allows us to consider their existence and uniqueness as subgradients of sublinear functions.

1.1 Definitions

Let $(\Omega, \mathcal{A}, \mu)$ be a measure space and $\mathcal{P}$ be a convex set of probability densities on Ω with respect to the measure μ. A random variable X takes values in Ω with unknown true density $p \in \mathcal{P}$. We refer to $\mathcal{P}$ and its elements as a *prediction set* and *predictive densities* for X, respectively. By $\mathcal{L}(\mathcal{P})$ we denote the set of all μ-measurable functions $f : \Omega \to \mathbb{R}$ such that

$$\int_\Omega |f(x)|\, p(x) d\mu(x) < \infty$$

for all $p \in \mathcal{P}$. We call the elements of $\mathcal{L}(\mathcal{P})$ *$\mathcal{P}$-integrable* functions.

A *scoring rule* $S : \mathcal{P} \to \mathcal{L}(\mathcal{P})$ assigns for each predictive density $q \in \mathcal{P}$ a $\mathcal{P}$-integrable function $S(q)$. The value of $S(q)$ at $x \in \Omega$ is interpreted as a numerical score assigned to the outcome x. We take scoring rules to be *positively orientated*, that is, they are viewed as incentives which a forecaster wishes to maximise. It is customary to term S *proper* if the expected value of S at q,

$$p \cdot S(q) := \int_\Omega S(q)(x)p(x)d\mu(x),$$

is maximised in q at the true density $q = p$, and *strictly proper*, if the true density is the only maximiser.

Strictly proper scoring rules could be used as a bonus system under which truth-telling is the only optimal long-term strategy (Gneiting and Raftery, 2007). For such an S, the optimal expected reward is the *(negative) entropy* induced by S,

$$\Phi : \mathcal{P} \to \mathbb{R}, \quad \Phi(p) = p \cdot S(p),$$

(Parry et al., 2012). In what follows, we refer to Φ simply as the entropy function associated to S, as there is no danger of confusion between negative and positive entropy functions in the present context. The *regret* for quoting q instead of the true density p is expressed by the function

$$D : \mathcal{P} \times \mathcal{P} \to \mathbb{R}, \quad D(p, q) = p \cdot S(p) - p \cdot S(q),$$

which in the statistics literature is also known as the *divergence* induced by S. In the present paper, we shall use the notions of entropy and divergence in a more general sense by replacing strict propriety with propriety.

General overviews of proper scoring rules may be found in Gneiting and Raftery (2007); Gneiting and Katzfuss (2014) in connection to probabilistic forecasting, and also in Dawid and Musio (2014), where the emphasis is on statistical inference. Theoretical aspects of proper scoring rules are studied in Dawid (2007); Grünwald and Dawid (2004); Williamson (2014). Frongillo and Kash (2014) investigate proper scoring rules in connection with the elicitation of private information. The remaining references throughout the text provide links to more specific uses of scoring rules.

1.2 Motivation and Scope of the Paper

In this paper we adopt the theoretical framework of Hendrickson and Buehler (1971). This approach is characterised by exploiting a beautiful connection with *Euler's homogeneous function theorem*, which presupposes that we extend our quantities of interest as homogeneous functions to the conic hull of the prediction set. To that end, we introduce the *prediction cone* $\mathcal{P}^+ = \{\lambda p \,|\, \lambda > 0, \, p \in \mathcal{P}\}$ and extend S and Φ to $\mathcal{P}^+$ as homogeneous functions of degrees zero and one, respectively. Any $\mathcal{P}$-integrable function q^* satisfying

$$\Phi(p) \geq p \cdot q^*, \quad \forall p \in \mathcal{P}^+,$$

with equality for $p = q$, is called a $\mathcal{P}$-*integrable subgradient* of Φ at q. The subgradient is called *strict* if the above inequality is strict for all $p \in \mathcal{P}^+$ not positively collinear to q. Suppose that Φ has a subgradient $S(q) \in \mathcal{L}(\mathcal{P})$ at each $q \in \mathcal{P}^+$ and the resulting

map $S : \mathcal{P}^+ \to \mathcal{L}(\mathcal{P})$ is homogeneous of degree zero. We call S a $\mathcal{P}$-*integrable subgradient* of Φ on $\mathcal{P}^+$. We recall that a (strictly) convex homogeneous function of degree one is a *(strictly) sublinear function*. We may now state Hendrickson and Buehler's classical result in a slightly more contemporary language.

Theorem 1.1 *Let $\mathcal{P}$ be a prediction set with respect to the measure space $(\Omega, \mathcal{A}, \mu)$. A scoring rule $S : \mathcal{P}^+ \to \mathcal{L}(\mathcal{P})$ is (strictly) proper if and only if there is a (strictly) sublinear function $\Phi : \mathcal{P}^+ \to \mathbb{R}$ such that S is a subgradient of Φ on $\mathcal{P}^+$.*

Theorem 1.1 provides us with a basic but insufficient theoretical framework to discuss the titular question of this paper. In support of this claim, in Example B.2 we show the existence of a sublinear function that has unique but non-$\mathcal{P}$-integrable subgradients at some points of its domain, while at other points it has multiple $\mathcal{P}$-integrable subgradients. The most important structure missing in Theorem 1.1 is the notion of *interior* of a convex domain, which lies at the intersection of geometry, algebra, and topology, and may have different incarnations depending on the context (Borwein and Vanderwerff, 2010; Rockafellar, 1972). For example, studying proper local scoring rules on discrete sample spaces, Dawid et al. (2012) apply Theorem 1.1 in a context where the prediction cone is the interior of the positive orthant in $\mathbb{R}^d$. In this case, well-known results from convex analysis give necessary and sufficient conditions for an affirmative answer to our basic question. The real focus of our paper is thus the non-Euclidean case in the abstract measure-theoretic setting introduced above.

In Proposition 2.4 and Example B.3, we show that at boundary points sublinear functions have either no subgradient, or infinitely many. Therefore, it is paramount to try to define entropy functions on interiors of positive cones. In infinite dimensions, however, this is not always possible. Indeed, it is well-known that the positive cones in many natural function spaces (such as the Lebesgue L^p-spaces over $\mathbb{R}^d$) have empty interiors (Borwein and Lewis, 1992) and are negligible sets in terms of Baire category. This calls for a more subtle approach to our problem in which we need to refine our notion of interior and boundary. Inspired by geometric functional analysis, we adapt an algebraic refinement of the notion of interior of convex sets, whose better known topological analogues are often referred to as *quasi-interior* (Fullerton and Braunschweiger, 1963; Borwein and Lewis, 1992). Common entropies whose domains are positive cones with empty interior but nonempty quasi-interior are the *Shannon entropy*, the *Hyvärinen entropy*, and in principle, the entropies associated with the *proper local scoring rules of arbitrary orders*. These entropies are formally not differentiable functions but possess directional derivatives on large subspaces, which display similar properties to standard gradients.

Other entropies, such as those that are associated with the families of *power scoring rules* and *pseudospherical scoring rules* may be extended continuously to open cones in normed spaces that contain signed densities. Geometrically, this setting is similar to the Euclidean setting. One applies the supporting hyperplane theorem and other standard results in analysis relating subgradients and Gâteaux derivatives. The latter entropies are Gâteaux differentiable (either everywhere or outside a negligible set), which we illustrate in the context of the *quadratic scoring rule*.

The original part of the paper is concerned with the analysis of the notion of $\mathcal{P}$-integrable subgradient introduced by Hendrickson and Buehler (1971) and the associated most ba-

sic general framework for proper scoring rules. To address the question of existence and uniqueness of proper scoring rules, we equip this framework with a notion of algebraic quasi-interior. As an illustration, we show that the Hyvärinen scoring rule is the unique 0-homogeneous $\mathcal{P}$-integrable subgradient of its entropy function on the (non-empty) quasi-interior of a suitable positive cone.

The paper is organised as follows. In Section 2, we introduce the notation and present all the background facts. Section 3 contains our main results which formulate necessary and sufficient conditions for existence and uniqueness of subgradients of entropy functions. In Section 4, we illustrate the theory with applications to three prototypical entropy functions, namely, the Shannon, Hyvärinen, and quadratic entropy. These examples formalise the meaning with respect to which we may consider each entropy to have a uniquely associated proper scoring rule. We complete the main part of the paper in Section 5 with some closing remarks. The proofs of all formal assertions made in the text are given in Appendix A. In Appendix B, we present additional facts that illustrate various points made in the Introduction or later in the text.

2. Notation and Preliminaries

Let E, E_1, E_2 be sets of μ-measurable functions on Ω. For $\alpha \in \mathbb{R}$, we use the notation

$$\alpha E_1 = \{\alpha f \,|\, f \in E_1\}$$
$$E_1 + E_2 = \{f + g \,|\, f \in E_1, g \in E_2\}.$$

The (*blunt*) *cone* of E is the set $E^+ = \{\lambda f \,|\, \lambda > 0, f \in E\}$, while the *pointed cone* of E is the set $E^+ \cup \{0\}$. The *convex hull* of E,

$$\operatorname{co} E = \left\{ \sum_{i=1}^{k} \alpha_i f_i \,\middle|\, k \geq 1,\ f_i \in E,\ \alpha_i \geq 0,\ \sum_{i=1}^{k} \alpha_i = 1 \right\},$$

is the set of all convex combinations of elements of E. The *conic hull* of E,

$$\operatorname{cone} E = \left\{ \sum_{i=1}^{k} \alpha_i f_i \,\middle|\, k \geq 1,\ f_i \in E,\ \alpha_i \geq 0 \right\},$$

is the set of all conic combinations of elements of E. By

$$\operatorname{span} E = \left\{ \sum_{i=1}^{k} \alpha_i f_i \,\middle|\, k \geq 1,\ f_i \in E,\ \alpha_i \in \mathbb{R} \right\}$$

we denote the set of all linear combinations of elements of E, and we refer to it as the *linear span* of E.

A set E is called *convex* if $\operatorname{co} E = E$, a *cone* if $E = E^+$ or $E = E^+ \cup \{0\}$, a *convex cone* if $E = \operatorname{cone} E$ or $E = \operatorname{cone} E \setminus \{0\}$, and a *linear space* if $E = \operatorname{span} E$. If E is convex, $E^+ = \operatorname{cone} E \setminus \{0\}$ is a convex cone.

The *epigraph* of $\Phi : E \to \mathbb{R}$ is the set in $\operatorname{span} E \times \mathbb{R}$ given by

$$\operatorname{epi} \Phi = \{(f, y) \,|\, f \in E, y \in \mathbb{R}, y \geq \Phi(f)\}.$$

The *graph* of Φ is the set $\{(f, \Phi(f)) \mid f \in E\}$.

A function $\Phi : E \to \mathbb{R}$ is called *convex* if its epigraph is a convex set. The definition implies that E is convex. Therefore, Φ is convex if, for any $f, g \in E$ and $\lambda \in (0, 1)$, Φ satisfies

$$\Phi((1 - \lambda)f + \lambda g) \leq (1 - \lambda)\Phi(f) + \lambda\Phi(g).$$

If the inequality is strict for $f \neq g$, then Φ is called *strictly convex.*

A function $\Phi : E^+ \to \mathbb{R}$ is said to be *(positively) homogeneous of degree k*, for $k \in \mathbb{R}$, or *(positively) k-homogeneous*, if for every $f \in E^+$ and every $\lambda > 0$, it holds $\Phi(\lambda f) = \lambda^k \Phi(f)$. A function $\Phi : E \to \mathbb{R}$ is said to be *subadditive* if Φ satisfies

$$\Phi(f + g) \leq \Phi(f) + \Phi(g)$$

for all $f, g \in E$, and *strictly subadditive*, if the above inequality is strict for $f \neq g$. We need to modify slightly the latter definition in the case when $\Phi : E^+ \to \mathbb{R}$ is 1-homogeneous. Then we say that Φ is *strictly subadditive* if the above inequality is strict whenever $f, g \in E^+$ are not positively collinear. Functions that are 1-homogeneous and (strictly) subadditive are called *(strictly) sublinear.* It is easy to see that $\Phi : E^+ \to \mathbb{R}$ is (strictly) sublinear if and only if Φ is (strictly) convex on E and 1-homogeneous on E^+.

Let $\mathcal{P}$ be a prediction set with respect to $(\Omega, \mathcal{A}, \mu)$ and let $E \subset \text{span}\,\mathcal{P}$. By $E^\perp$ we denote the *annihilator* of E in $\mathcal{L}(\mathcal{P})$, that is, all $f \in \mathcal{L}(\mathcal{P})$ such that

$$p \cdot f = 0$$

for all $p \in E$. Clearly, $E^\perp$ is a linear subspace of $\mathcal{L}(\mathcal{P})$. In the case when $E^\perp = \{0\}$, we say that E has a *trivial annihilator.*

By a *direction* in a vector space we understand the equivalence class of all positively collinear vectors to a given nonzero vector. Note that any 0-homogeneous function is a function of directions. For $q \in \mathcal{P}^+$, we define the *set of directions* from q to the points in $\mathcal{P}^+$ as

$$\mathcal{D}(q) = \{p \in \text{span}\,\mathcal{P} \mid \exists \epsilon_p > 0, \, \forall t \in [0, \epsilon_p], \, q + tp \in \mathcal{P}^+\}$$
$$= \{p \in \text{span}\,\mathcal{P} \mid \exists \epsilon_p > 0, \, q + \epsilon_p p \in \mathcal{P}^+\}.$$

We have the latter identity due to the convexity of $\mathcal{P}^+$.

A point $q \in \mathcal{P}^+$ is an *algebraically interior* point of $\mathcal{P}^+$ if $\mathcal{D}(q) = \text{span}\,\mathcal{P}$. The collection of all algebraically interior points of $\mathcal{P}^+$ is called the *algebraic interior* of $\mathcal{P}^+$. In the case of a topological vector space, the topological interior of a set is always contained in the algebraic interior of the set. Moreover, when the topological interior is not empty, the two notions coincide. If q is not algebraically interior for $\mathcal{P}^+$, that is, $\mathcal{D}(q) \neq \text{span}\,\mathcal{P}$, we say that q is a *boundary point* for $\mathcal{P}^+$. If $\mathcal{P}^+$ has empty algebraic interior, then the prediction cone consists entirely of boundary points. This case occurs frequently in the context of continuous sample spaces, see e.g. Proposition B.1.

Lemma 2.1 *For each $q \in \mathcal{P}^+$, we have the representation*

$$\mathcal{D}(q) = \text{cone}(\mathcal{P}^+ - q).$$

For a point $q \in \mathcal{P}^+$, we define $\mathcal{O}(q) = \mathcal{D}(q) \cap -\mathcal{D}(q)$. This is the subset of directions in $\mathcal{D}(q)$ whose inverse is also in $\mathcal{D}(q)$. The set may be identified with these directions in span $\mathcal{P}$ along which there is an open line segment that contains q and is contained in $\mathcal{P}^+$. Clearly, q is algebraically interior for $\mathcal{P}^+$ if and only if $\mathcal{O}(q) = \mathcal{D}(q) = \text{span} \, \mathcal{P}$. By construction, $\mathcal{O}(q)$ is a linear subspace of span $\mathcal{P}$. The sets of directions $\mathcal{D}(q)$ and $\mathcal{O}(q)$ are instrumental for defining various notions of directional derivatives.

The most basic directional derivative is the following one.

Definition 2.2 *For a function $\Phi : \mathcal{P}^+ \to \mathbb{R}$, the* right directional derivative *of Φ at $q \in \mathcal{P}^+$ along $p \in \mathcal{D}(q)$ is defined as*

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

if the limit exists.

We gather below the main properties of $\Phi'_+(p, q)$.

Proposition 2.3 *Let $\Phi : \mathcal{P}^+ \to \mathbb{R}$ be a sublinear function and $q \in \mathcal{P}^+$. We have*

(a) for each $p \in \mathcal{D}(q)$,

$$\Phi'_+(p, q) = \inf_{t > 0} \frac{\Phi(q + tp) - \Phi(q)}{t} \in \mathbb{R} \cup \{-\infty\},$$

and the infimum is finite for $p \in \mathcal{O}(q)$;

(b) $\Phi'_+(\cdot, q) : \mathcal{D}(q) \to \mathbb{R} \cup \{-\infty\}$ is sublinear;

(c) for each $\lambda > 0$, $\Phi'_+(p, \lambda q) = \Phi'_+(p, q)$;

(d) for each $p \in \mathcal{P}^+$,

$$\Phi(p) \geq \Phi'_+(p, q),$$

with equality for $p = q$;

(e) for each $p \in \mathcal{O}(q)$, $-\Phi'_+(-p, q) \leq \Phi'_+(p, q)$;

(f) the set

$$\mathcal{O}'(q) = \{p \in \mathcal{O}(q) \mid -\Phi'_+(-p, q) = \Phi'_+(p, q)\}$$

is a linear subspace of $\mathcal{O}(q)$ and the restriction $\Phi'_+(\cdot, q)\big|_{\mathcal{O}'(q)}$ is linear.

We next consider the other two types of directional derivatives. First, if we take the limit (1) with the restriction $t \leq 0$ instead $t \geq 0$, we obtain the *left directional derivative* of Φ, denoted $\Phi'_-(\cdot, q)$. It is easy to see that $\Phi'_-(\cdot, q)$ can be defined on $\mathcal{O}(q)$ and we have $\Phi'_-(p, q) = -\Phi'_+(-p, q)$, for each $p \in \mathcal{O}(q)$. Thus part (e) above can be rewritten as

$$\Phi'_-(p, q) \leq \Phi'_+(p, q)$$

for all $p \in \mathcal{O}(q)$. On the subspace $\mathcal{O}'(q)$ introduced above in part (f), we have that

$$\Phi'_-(\cdot, q) = \Phi'_+(\cdot, q)$$

is in fact the *two-sided directional derivative* of Φ at q, denoted $\Phi'(\cdot, q)$. The latter can be defined as the limit (1) without any restriction on t. In the most important case in practice, we have that $\mathcal{O}(q) = \mathcal{O}'(q)$. If in addition $\mathcal{O}(q) \neq \operatorname{span} \mathcal{P}$, then Φ has no standard functional derivative. For an illustration of this fact in the context of Shannon and Hyvärinen entropies, see Section 4.

By $\operatorname{Lin} \mathcal{P}$ we denote the space of all real-valued linear functionals on $\operatorname{span} \mathcal{P}$, i.e., the algebraic dual of $\operatorname{span} \mathcal{P}$. By "$\cdot$" we denote the bilinear pairing on $\operatorname{span} \mathcal{P} \times \operatorname{Lin} \mathcal{P}$, so if $q \in \operatorname{span} \mathcal{P}$ and $q^* \in \operatorname{Lin} \mathcal{P}$, $q \cdot q^*$ is the value of q^* at q.

Let $\Phi : \mathcal{P}^+ \to \mathbb{R}$ be 1-homogeneous. We say that $q^* \in \operatorname{Lin} \mathcal{P}$ is a *subgradient* of Φ at q if

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

for all $p \in \mathcal{P}^+$, with equality for $p = q$. The collection of all subgradients of Φ at q is called the *subdifferential* of Φ at q and is denoted by $\partial \Phi(q)$. A subgradient q^* is *strict* if and only if the inequality $\Phi(p) > p \cdot q^*$ holds for all $p \in \mathcal{P}^+$ not positively collinear with q.

If $h \in \operatorname{Lin} \mathcal{P}$, the *hyperplane H* in $\operatorname{span} \mathcal{P} \times \mathbb{R}$ given by

$$z = p \cdot h, \quad \forall p \in \operatorname{span} \mathcal{P},$$

supports Φ at q if the epigraph of Φ lies above H, and H contains the point $(q, \Phi(q))$. Clearly, H supports Φ at q if and only if $h \in \partial \Phi(q)$.

The following proposition describes the intimate connection between one-sided and two-sided directional derivatives and the subdifferential of a sublinear function.

Proposition 2.4 *For a point $q \in \mathcal{P}^+$, we have*

(a) $q^* \in \partial \Phi(q)$ if and only if

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

for all $p \in \mathcal{P}^+$, with equality for $p = q$;

(b) if $\mathcal{D}(q) = \operatorname{span} \mathcal{P}$ and $\Phi'(\cdot, q)$ exists on $\operatorname{span} \mathcal{P}$, then $\partial \Phi(q) = \{\Phi'(\cdot, q)\}$;

(c) if $\mathcal{D}(q) = \operatorname{span} \mathcal{P}$ and $\Phi'(\cdot, q)$ does not exist on $\operatorname{span} \mathcal{P}$, then $\partial \Phi(q)$ has multiple elements;

(d) if $\mathcal{D}(q) \neq \operatorname{span} \mathcal{P}$ and $\Phi'_+(p, q)$ is finite for all $p \in \mathcal{P}^+$, then $\partial \Phi(q)$ has multiple elements;

(e) if $\mathcal{D}(q) \neq \operatorname{span} \mathcal{P}$ and there is $p \in \mathcal{P}^+$ such that $\Phi'_+(p, q) = -\infty$, then $\partial \Phi(q) = \emptyset$.

Part (a) above is the standard characterisation of the subdifferential of a sublinear function. Parts (b) and (c) give additional information in the case of algebraically interior points. Parts (d) and (e) do the same for boundary points. Notice that the latter imply the statement from the Introduction that at boundary points either the existence or uniqueness of subgradient fails. (See also Example B.3.) In the next section, we show that uniqueness might be sometimes recovered at certain boundary points if we confine ourselves to a regularity class such as $\mathcal{L}(\mathcal{P})$.

We next give a formal definition of a scoring rule and elaborate some of its implications.

Definition 2.5 *Let $\mathcal{P}$ be a prediction set with respect to the measure space $(\Omega, \mathcal{A}, \mu)$. Any 0-homogeneous map $S : \mathcal{P}^+ \to \mathcal{L}(\mathcal{P})$ is called a* scoring rule.

If X is a random variable on Ω with unknown true density $p \in \mathcal{P}$, then for each predictive density $q \in \mathcal{P}^+$, $S(q)(X)$ is a random function of X. The condition $S(q) \in \mathcal{L}(\mathcal{P})$ guarantees that the expectation of S is always finite. The *uncertainty function* associated to S is the function $\Phi : \mathcal{P}^+ \to \mathbb{R}$, $\Phi(p) = p \cdot S(p)$. Clearly, Φ is 1-homogeneous. When S is proper, it is customary to call Φ an *entropy function*.

Suppose now that $S : \mathcal{P}^+ \to \mathcal{L}(\mathcal{P})$ is a proper scoring rule with entropy Φ. The condition that the expected score of S is maximised in q at the true density $q = p$ means that S satisfies the inequality

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

for each $p, q \in \mathcal{P}^+$, with equality for $q = p$. If S is strictly proper, then p is the only maximiser up to a scaling factor. In this case, the inequality above is strict for any q that is not positively collinear to p. So, the assumption of propriety is equivalent to S being a subgradient of Φ on $\mathcal{P}^+$. Moreover, strict propriety corresponds to strict subgradients on $\mathcal{P}^+$. The existence of a subgradient on $\mathcal{P}^+$ implies that Φ is sublinear, see Lemma A.1. We conclude that (strictly) proper scoring rules are $\mathcal{P}$-integrable subgradients of (strictly) sublinear functions. Therefore, it is reasonable in the context of scoring rules to restrict the notion of subgradient to the class $\mathcal{L}(\mathcal{P}) \subset \mathrm{Lin}(\mathcal{P})$. In the next section, and in particular in Theorem 3.1 and Theorem 3.2, we discuss the existence and uniqueness of $\mathcal{P}$-integrable subgradients.

In some special cases, we may add to our notion of subgradient a topological structure. Let $\mathcal{P}^+$ be a prediction cone such that $\mathrm{span}\,\mathcal{P}$ may be identified with a normed space $(N, \|\cdot\|)$, and let the continuous dual of N, denoted N^*, be a subset of $\mathcal{L}(\mathcal{P})$. Suppose that $\mathcal{P}^+ \subset \mathcal{C}$, where $\mathcal{C}$ is an open convex cone in N, and Φ may be extended to $\mathcal{C}$ as a continuous sublinear function.

We recall that Φ is *Gâteaux differentiable* at $q \in \mathcal{C}$ if there is $q^* \in N^*$ such that for every $p \in N$, 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 Φ at q and is also denoted by $\nabla \Phi(q)$. Notice that by definition the Gâteaux derivative is applicable only to interior points. See Theorems 3.3 and 3.4 for an answer to our two main questions.

If Φ is Gâteaux differentiable at q, taking $p = q$ in the above limit, we recover *Euler's homogeneous function* theorem

$$q \cdot \nabla \Phi(q) = \Phi(q).$$

More generally, if Φ is sublinear and has a subgradient S on $\mathcal{P}^+$, then we have that $q \cdot S(q) = \Phi(q)$, for every $q \in \mathcal{P}^+$, (Hendrickson and Buehler, 1971). The proof also follows from Proposition 2.4 (a) and Proposition 2.3 (d). This beautiful generalisation of Euler's theorem is only visible after extending S and Φ to denormalised densities as homogeneous functions.

Suppose now that a scoring rule $S : \mathcal{P} \to \mathcal{L}(\mathcal{P})$ is given. Then, setting

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

for any $q \in \mathcal{P}^+$, extends S as a 0-homogeneous function to the prediction cone. Here

$$q \cdot 1 = \int_\Omega q(x) d\mu(x)$$

is the normalising constant of q. Similarly, let an entropy function $\Phi : \mathcal{P} \to \mathbb{R}$ be given. Setting

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

for any $q \in \mathcal{P}^+$, extends Φ as a 1-homogeneous function to the prediction cone. See Section 4 for an illustration. Working directly with denormalised predictive densities could also be advantageous in numerical computation (Hyvärinen, 2005, 2007; Dawid and Musio, 2012, 2014).

3. Main Results

Our first result gives a necessary and sufficient condition for existence of a $\mathcal{P}$-integrable subgradient at a point. The result can be easily generalised to subgradients on $\mathcal{P}^+$.

Theorem 3.1 *Let $\Phi : \mathcal{P}^+ \to \mathbb{R}$ be a sublinear function. Then Φ has a $\mathcal{P}$-integrable subgradient at a point $q \in \mathcal{P}^+$ if and only if there is $q^* \in \mathcal{L}(\mathcal{P})$ such that*

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

for all $p \in \mathcal{P}^+$, with equality for $p = q$.

In the light of Theorem 1.1 and the above result, we call any sublinear function Φ an *entropy* if Φ has a $\mathcal{P}$-integrable subgradient at each point of its domain. In most cases of practical interest, one may choose the prediction cone appropriately so that $\Phi'_+(\cdot, q) = q^*$ for some $q^* \in \mathcal{L}(\mathcal{P})$. This means that $\Phi'_+(\cdot, q)$ is a $\mathcal{P}$-integrable subgradient of Φ at q and that $\Phi'_+(\cdot, q) = \Phi'(\cdot, q)$ is also a two-sided directional derivative on the subspace $\mathcal{O}(q)$ of span $\mathcal{P}$. In our next result, we show that if $\mathcal{O}(q)$ is a sufficiently large subspace, then $\Phi'_+(\cdot, q)$ is the unique $\mathcal{P}$-integrable subgradient of Φ at q.

Theorem 3.2 *Let $\mathcal{P}$ be a prediction set and $\Phi : \mathcal{P}^+ \to \mathbb{R}$ be a sublinear function. Suppose that at a point $q \in \mathcal{P}^+$ the subspace $\mathcal{O}(q)$ of span $\mathcal{P}$ has a trivial annihilator in $\mathcal{L}(\mathcal{P})$. If there is a $q^* \in \mathcal{L}(\mathcal{P})$ such that*

$$p \cdot q^* = \Phi'_+(p, q) \tag{2}$$

for all $p \in \mathcal{P}^+$, then q^* is the unique $\mathcal{P}$-integrable subgradient of Φ at q.

In the above result, the condition that $\mathcal{O}(q)$ has a trivial annihilator in $\mathcal{L}(\mathcal{P})$ can be interpreted to say that the set of directions at which $q \in \mathcal{P}^+$ is boundary to the cone $\mathcal{P}^+$ is negligible. The latter condition represents an algebraic analogue to the property of q being a *quasi-interior point* of $\mathcal{P}^+$, which is better known in its topological forms presented in Fullerton and Braunschweiger (1963); Borwein and Lewis (1992). The collection of all quasi-interior points of $\mathcal{P}^+$ is the *quasi-interior* of $\mathcal{P}^+$. As an illustration, in the next section we define Shannon and Hyvärinen entropies on positive cones with nonempty quasi-interiors.

Presently, however, we do not investigate the proposed variant of quasi-interior in full. This analysis is not necessary for the application of Theorem 3.2 and may be a subject of future work. Notice also that uniqueness of subgradient is understood and valid only within the class $\mathcal{L}(\mathcal{P})$.

We now consider the case of topological subgradients. Our main assumption is the following:

$$\begin{cases} \mathcal{P}^+ \subset \mathcal{C}, \text{ where } \mathcal{C} \text{ is an open convex cone in a normed space } N \\ \Phi : \mathcal{C} \to \mathbb{R} \text{ is a continuous sublinear function.} \end{cases} \tag{3}$$

Theorem 3.3 *If* (3) *holds, then* Φ *admits a subgradient* $S : \mathcal{C} \to N^*$.

The result is generally known as the *supporting hyperplane theorem*. For proof see e.g. Niculescu and Persson (2006); Borwein and Vanderwerff (2010); Zalinescu (2002); Rudin (1973). Any subgradient $S : \mathcal{C} \to N^*$ of Φ may be identified with a proper scoring rule on $\mathcal{P}^+$ by restricting S to $\mathcal{P}^+$.

Theorem 3.4 *Assume* (3). *Then,* Φ *is Gâteaux differentiable on* $\mathcal{C}$ *if and only if* Φ *admits a unique subgradient* $S : \mathcal{C} \to N^*$. *In this case* $S = \nabla\Phi$ *is the Gâteaux derivative of* Φ.

This is a standard result in convex analysis. See e.g. Borwein and Vanderwerff (2010); Zalinescu (2002). See Example B.2 for an illustration of the case where the assumption $N^* \subset \mathcal{L}(\mathcal{P})$ is not satisfied.

4. Applications

In this section, we apply our main results to three important entropies: *Shannon entropy*, *Hyvärinen entropy*, and *quadratic entropy*. For each entropy, we investigate an appropriate domain with nonempty quasi-interior for which we show the existence of a unique subgradient.

4.1 Shannon Entropy

The *Shannon entropy function* for densities on $\mathbb{R}^d$ is given by

$$\Phi(p) = \int_{\mathbb{R}^d} p(x) \ln \frac{p(x)}{p \cdot 1} dx \tag{4}$$

where $p(x) \geq 0$ is assumed to be sufficiently regular. More facts about Shannon entropy may be found e.g. in Dawid (2007); Parry et al. (2012); Dawid et al. (2012).

We first show that Shannon entropy may only be defined for nonnegative functions in a natural way. The kernel of Φ is the function $\phi(t) = t \ln t$ for $t > 0$ and $\phi(0) = 0$. Clearly, $\phi(t)$ is strictly convex on $t \geq 0$ since, for $t > 0$, $\phi''(t) = 1/t > 0$, and ϕ is continuous at the endpoint $t = 0$. Notice that $\phi(t)$ has a vertical tangent at $t = 0$ since $\phi'(t) = \ln t + 1$. We conclude that $\phi(t)$ cannot be extended as a convex function to $t < 0$. This furnishes our claim.

The positive cone of $L^1(\mathbb{R}^d)$ comprises of all nonnegative functions in $L^1(\mathbb{R}^d)$ and is denoted by $L^1_+(\mathbb{R}^d)$. In Proposition B.1 we give a direct proof that $L^1_+(\mathbb{R}^d)$ is a nowhere

dense subset of $L^1(\mathbb{R}^d)$. Since the domain of Shannon entropy is a subset of $L^1_+(\mathbb{R}^d)$, it too is a nowhere dense set.

We now proceed to find a suitable prediction set. For $a \geq d+1$, we set

$$\mathcal{P}^+ = \left\{ p \in C(\mathbb{R}) \,\Big|\, p(x) > 0 \,, \exists C_1, C_2 > 0 \,:\, \frac{C_1}{(1+|x|)^a} \leq p(x) \leq \frac{C_2}{(1+|x|)^{d+1}} \right\}.$$

Notice that $\mathcal{L}(\mathcal{P}) \subset L^1_{\mathrm{loc}}(\mathbb{R}^d)$. Indeed, for any $f \in \mathcal{L}(\mathcal{P})$ consider

$$p_t(x) = \begin{cases} 1 & 0 < |x| < t \\ \left(\frac{1+t}{1+|x|}\right)^{d+1} & t \leq |x|. \end{cases}$$

Since $p_t \in \mathcal{P}^+$, the $\mathcal{P}$-integrability of f implies that

$$\int_{|x| \leq t} |f(x)|\, dx < \infty$$

for all $t > 0$.

Let us next see that for any $q \in \mathcal{P}^+$, $\mathcal{O}(q)$ has a trivial annihilator in $\mathcal{L}(\mathcal{P})$. Clearly, $\mathcal{O}(q)$ contains all $p \in \operatorname{span} \mathcal{P}$ that have faster or equal decay at infinity compared to q. Suppose that $f \in \mathcal{O}(q)^\perp$. Choosing an appropriate approximation of the identity, $\{p_n\}$, $p_n \in \mathcal{O}(q)$, we get that $f * p_n(x) \to f(x)$ for every x in the Lebesgue set of f. Hence $f = 0$ a.e. on $\mathbb{R}^d$. We conclude that $\mathcal{O}(q)^\perp = \{0\}$.

After this preparation, we may now define Φ rigorously as the map from $\mathcal{P}^+$ to $\mathbb{R}$ given by (4). Strict convexity of Φ follows from the strict convexity of $t \ln t$, for $t \geq 0$, while its 1-homogeneity is trivial. Therefore, Φ is strictly sublinear on $\mathcal{P}^+$. Let us compute the right directional derivative of Φ.

For $q \in \mathcal{P}^+$ and $p \in \mathcal{D}(q)$, we set $q_t = q + tp$. We have

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

$$= p \cdot \ln \frac{q}{q \cdot 1} + q \cdot \left(\frac{p}{q} - \frac{p \cdot 1}{q \cdot 1} \right)$$

$$= p \cdot \ln \frac{q}{q \cdot 1}.$$

Therefore,

$$\Phi'_+(p, q) = \int_{\mathbb{R}^d} p(x) \ln \frac{q(x)}{q \cdot 1}\, dx.$$

Clearly, the function

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

is in $\mathcal{L}(\mathcal{P})$. Indeed, the claim follows from the fact that $S(q)$ is continuous in x and grows logarithmically as $|x| \to \infty$. In view of Theorem 3.2, S is the unique $\mathcal{P}$-integrable subgradient of Φ on $\mathcal{P}^+$ since $\Phi'_+(p, q) = p \cdot S(q)$ for every $p, q \in \mathcal{P}^+$. The map is known as the *logarithmic scoring rule*.

The uniqueness of the logarithmic scoring rule as a subgradient of Shannon entropy is in no way an absolute fact. Using the Hahn-Banach theorem as illustrated in Example B.3 and the fact that $L_+^1(\mathbb{R}^d)$ consists entirely of boundary points, one may construct other subgradients of Φ that lie outside $\mathcal{L}(\mathcal{P})$. Moreover, if q lies on the *quasi-boundary* of $\mathcal{P}^+$ (i.e. the points where the condition $\mathcal{O}(q)^\perp = \{0\}$ is violated), then uniqueness will fail even within $\mathcal{L}(\mathcal{P})$.

4.2 Hyvärinen Entropy

Hyvärinen entropy for densities on $\mathbb{R}^d$ is defined as

$$\Phi(p) = \int_{\mathbb{R}^d} \frac{|\nabla p(x)|^2}{p(x)} dx. \tag{5}$$

Here ∇ is the gradient on $\mathbb{R}^d$. Hyvärinen and related entropies are considered e.g. in Parry et al. (2012); Ehm and Gneiting (2012); Forbes and Lauritzen (2014); Dawid and Musio (2012); Hyvärinen (2005, 2007); Sánchez-Moreno et al. (2012).

We first show that there is no natural way to extend Hyvärinen entropy to signed densities. For simplicity, we confine ourselves to the case $d = 1$. Suppose that p changes sign at some $x_0 \in \mathbb{R}$ that has multiplicity one. The assumption is generic and it means that x_0 is not an inflection point of p. It follows that the above integral is divergent at x_0. Indeed, the claim is a direct consequence of the asymptotic expansion of the term

$$\frac{|p'(x)|^2}{p(x)} = \frac{1}{x - x_0} + O(x - x_0)$$

near x_0. On the other hand, if p has a zero of higher multiplicity at x_0, one may check that the above asymptotics will be bounded and the integral will be convergent in a neighbourhood of x_0. Nevertheless, the example shows that Φ cannot be generally defined for densities that change sign.

We proceed to define a suitable domain for Φ. Suppose that $\mathcal{P}^+$ consists of all positive, twice continuously differentiable functions $p(x)$ on $\mathbb{R}^d$ that satisfy the bounds:

(a) there are $C_1 > 0$ and $k > 0$ such that

$$\left| \frac{\nabla p(x)}{p(x)} \right| + \left| \frac{\Delta p(x)}{p(x)} \right| \leq C_1(1 + |x|)^k;$$

(b) there is $C_2 > 0$ such that

$$|p(x)| \leq \frac{C_2}{(1 + |x|)^{d+1+k^2}},$$

where $\Delta = \partial^2/\partial x_1^2 + \cdots + \partial^2/\partial x_d^2$ is the *Laplacian* on $\mathbb{R}^d$. In view of the above, we have the following limit

$$\lim_{R \to \infty} \frac{1}{R} \int_{|y|=R} \left(\frac{y \nabla q(y)}{q(y)} \right) p(y) dy = 0 \tag{6}$$

for any $p, q \in \mathcal{P}^+$. Note that here

$$y \nabla q(y) = y_1 \frac{\partial q(y)}{\partial y_1} + \cdots + \frac{y_d \partial q(y)}{\partial y_d}$$

denotes the *scalar product* of y and $\nabla q(y)$ and the integral in (6) is a surface integral over the sphere centred at the origin of radius R. The class $\mathcal{P}$ is broad, e.g. it contains the Gaussians, and all positive continuous densities that have bounded first and second-order derivatives and decay at infinity sufficiently fast. Just like in Section 4.1, we have that $\mathcal{L}(\mathcal{P}) \subset L^1_{\mathrm{loc}}(\mathbb{R}^d)$ and that for any $q \in \mathcal{P}^+$ the annihilator of $\mathcal{O}(q)$ in $\mathcal{L}(\mathcal{P})$ is trivial. In the light of Proposition B.1, $\mathcal{P}^+$ is nowhere dense in $L^1(\mathbb{R}^d)$ as $\mathcal{P}^+ \subset L^1_+(\mathbb{R}^d)$.

We now formally define Hyvärinen entropy as the map from $\mathcal{P}^+$ to $\mathbb{R}$ given in (5). Convexity of Φ follows from the convexity of the function

$$\phi(t, t_1, \ldots, d_d) = \frac{t_1^2 + \cdots + t_d^2}{t}, \quad \text{for } t > 0, \ (t_1, \ldots, t_d) \in \mathbb{R}^d,$$

while its 1-homogeneity is trivial. Hence, Φ is sublinear. Let us compute its right directional derivative.

For $q \in \mathcal{P}^+$ and $p \in \mathcal{D}(q)$, we set $q_t = q + tp$. We have

$$\lim_{t \to 0^+} \frac{\Phi(q + tp) - \Phi(q)}{t} = \int_{\mathbb{R}^d} \frac{d}{dt}\Big|_{t=0} \left(\frac{|\nabla q_t(x)|^2}{q_t(x)} \right) dx$$

$$= \int_{\mathbb{R}^d} \left(2 \frac{\nabla q(x)}{q(x)} \frac{\nabla p(x)}{p(x)} - \frac{|\nabla q(x)|^2}{q^2(x)} \right) p(x)dx.$$

By integration by parts we get

$$\int_{|x| \leq R} \left(\frac{2 \nabla q(x) \nabla p(x)}{q(x)} - \frac{|\nabla q(x)|^2}{q^2(x)} p(x) \right) dx$$

$$= \int_{|x| \leq R} \left(-\frac{2 \Delta q(x)}{q(x)} + \frac{|\nabla q(x)|^2}{q^2(x)} \right) p(x)dx + \frac{2}{R} \int_{|y|=R} \left(\frac{y \nabla q(y)}{q(y)} \right) p(y)dy.$$

Letting $R \to \infty$ and using (6), we obtain

$$\Phi'_+(p, q) = \int_{\mathbb{R}^d} \left(-\frac{2 \Delta q(x)}{q(x)} + \frac{|\nabla q(x)|^2}{q^2(x)} \right) p(x)dx.$$

The assumptions on $\mathcal{P}^+$ guarantee that

$$S(q)(x) = -\frac{2 \Delta q(x)}{q(x)} + \frac{|\nabla q(x)|^2}{q^2(x)}$$

is $\mathcal{P}$-integrable for every $q \in \mathcal{P}^+$. In view of Theorem 3.2, $S(q)$ is the unique $\mathcal{P}$-integrable subgradient of Φ on $\mathcal{P}^+$. The map is known as the *Hÿvarinen scoring rule* (Parry et al., 2012).

In fact, $S(q)$ is a strict subgradient of Φ on $\mathcal{P}^+$. This can be shown if we notice that the divergence induced by S has the representation

$$p \cdot S(p) - p \cdot S(q) = \int_{\mathbb{R}^d} \left| \frac{\nabla p(x)}{p(x)} - \frac{\nabla q(x)}{q(x)} \right|^2 p(x) dx.$$

The latter identity can be proved by integration by parts. The divergence is zero if and only if

$$\nabla(\ln p(x) - \ln q(x)) = 0.$$

This is equivalent to $p = Cq$ for some constant $C > 0$, i.e., p and q being positively collinear. This concludes the proof of the claim.

4.3 Quadratic Entropy

Here we consider the *quadratic entropy*

$$\Phi(q) = \frac{1}{q \cdot 1} \int_\Omega q^2(x) dx, \tag{7}$$

where $(\Omega, \mathcal{A}, \mu)$ is a Lebesgue measure space with $\Omega \subset \mathbb{R}^d$. In what follows, we show that its Gâteaux derivative is the *quadratic scoring rule*, also known as *Brier score* (Brier, 1950). The quadratic entropy is a member of the important family of *power entropy functions*. The corresponding *power scoring rules* have been studied in connection to robust estimation e.g. in Basu et al. (1998); Kanamori and Fujisawa (2015, 2014).

We proceed to choose a suitable domain for Φ. In contrast to the previous two entropies we now introduce a topology. To that end, we begin with a description of some normed spaces. Let $w : \Omega \to [0, \infty)$ be a measurable function which we call a *weight*. By $L^p(\Omega, w)$, for $p \geq 1$, we denote the Lebesgue space of functions on Ω whose p-th power is absolutely integrable with respect to the weight $w(x)$. By $\|\cdot\|_{p,w}$ we denote the corresponding weighted L^p-norm. When w is identically equal to one we get the usual Lebesgue space and norm. In this case we drop w from our notation. We now set

$$w(x) = (1 + |x|)^{d+1}.$$

Notice that $L^2(\Omega, w)$ embeds continuously in $L^1(\Omega)$. Indeed, for $f \in L^1(\Omega)$, we have

$$\int_\Omega |f(x)|\, dx = \int_\Omega w^{-1/2}(x)\, |f(x)|\, w^{1/2}(x) dx$$
$$\leq \left(\int_\Omega w^{-1}(x) dx \right)^{\frac{1}{2}} \left(\int_\Omega |f(x)|^2\, w(x) dx \right)^{\frac{1}{2}}$$
$$\leq C \, \|f\|_{2,w},$$

where $C > 0$ is a constant. Clearly, $L^2(\Omega, w)$ also embeds continuously in $L^2(\Omega)$ and hence the same conclusion holds for $L^2(\Omega, w)$ for all intermediate spaces $L^p(\Omega)$ with $1 \leq p \leq 2$. Hence, we have the inequality

$$\|f\|_p \leq C \, \|f\|_{2,w}$$

for some fixed $C > 0$ and all $p \in [1, 2]$.

We have that $f \in L^2(\Omega, w)$ if and only if $fw^{1/2} \in L^2(\Omega)$. Clearly, the weight is needed only when Ω is unbounded as otherwise the weighted and the ordinary L^p-norms are equivalent. The continuous dual space of $L^2(\Omega, w)$ may be identified with the space $L^2(\Omega, w^{-1})$. Therefore, $g \in L^2(\Omega, w^{-1})$ if and only if $gw^{-1/2} \in L^2(\Omega)$. Hence, the dual space $L^2(\Omega, w^{-1})$ contains the constants and also the elements of $L^2(\Omega, w)$.

We now specify a prediction set $\mathcal{P} \subset L^2_+(\Omega, w)$ with the following property: there are constants $k_1 > 0$ and $k_2 > 0$ such that

$$k_1 \leq \|q\|_{2,w} \leq k_2$$

for all $q \in \mathcal{P}$. Choose $0 < \epsilon < \min(1, k_1)$. For $p \in L^2(\Omega)$, let $B_\rho(p)$ denote the open ball about p of radius $\rho > 0$. Choose $\delta > 0$ so small that for every $p \in B_\delta(0)$ we have $\|p\|_1 \leq \epsilon$ and $\|p\|_{2,w} \leq \epsilon$. Let $q \in \mathcal{P}$ and consider $r \in B_\delta(q)$. It is easy to show that

$$k_1 - \epsilon \leq \|r\|_{2,w} \leq k_2 + \epsilon$$

for all $r \in B_\delta(q)$. Similarly, we also have

$$1 - \epsilon \leq r \cdot 1 \leq 1 + \epsilon$$

for all $r \in B_\delta(q)$. Here we have used the fact that $r = p + q$, where $q \cdot 1 = 1$ and $\|p\|_1 \leq \epsilon$. We now set

$$\mathcal{C}_0 = \mathcal{P} + B_\delta(0) = \cup_{q \in \mathcal{P}} B_\delta(q).$$

It follows that $\mathcal{C}_0$ is convex as both $\mathcal{P}$ and $B_\delta(0)$ are convex. Finally, let $\mathcal{C} = \mathcal{C}_0^+$ be the cone of $\mathcal{C}_0$. Clearly, $\mathcal{C}$ is an open convex cone in $L^2(\Omega, w)$.

We may now formally define Φ as the map from $\mathcal{C}$ to $\mathbb{R}$ given by (7). We have that Φ is strictly convex on $\mathcal{C}_0$ as the kernel function $\phi(t) = t^2$ is strictly convex for $t \in \mathbb{R}$. Therefore, Φ is strictly sublinear on $\mathcal{C}$. It is not hard to see that Φ is also continuous on $\mathcal{C}$. Theorem 3.3 implies that Φ has a subgradient on $\mathcal{C}$. The following computation shows that Φ is Gâteaux differentiable. Indeed, for $q \in \mathcal{C}$ and $p \in L^2(\Omega, w)$, we have

$$\lim_{t \to 0} \frac{\Phi(q + tp) - \Phi(q)}{t} = \int_\Omega \frac{d}{dt}\bigg|_{t=0} \frac{(q(x) + tp(x))^2}{(q + tp) \cdot 1} dx$$

$$= 2 \int_\Omega \frac{q(x)p(x)}{q \cdot 1} dx - \int_\Omega \frac{q^2(x)}{(q \cdot 1)^2} dx \int_\Omega p(x) dx.$$

We obtain that

$$\nabla \Phi(q) = \frac{2q}{q \cdot 1} - \frac{q \cdot q}{(q \cdot 1)^2}$$

is the Gâteaux derivative of Φ as clearly $\nabla \Phi(q) \in L^2(\Omega, w^{-1})$. In view of Theorem 3.4, $S = \nabla \Phi|_{\mathcal{P}^+}$ defines a strictly proper scoring rule on $\mathcal{P}^+$. We have that $\nabla \Phi$ is the unique subgradient of quadratic entropy on the cone $\mathcal{C}$, but as discussed before, by using the Hahn-Banach theorem one may show that uniqueness fails on $\mathcal{P}^+$ when Ω is unbounded. The rule S is known as the *quadratic scoring rule*.

5. Conclusion

We were originally motivated to understand the implications of the fact that Shannon and Hyvärinen entropies are only finite on domains with empty interiors. As no notion of functional derivative is applicable to these entropies, the question whether the logarithmic and Hyvärinen scoring rules are the unique subgradients of their respective entropy functions is not obvious. In contrast, the quadratic entropy may be continuously extended to signed densities, which allows us to interpret the quadratic scoring rule as the Gâteaux derivative of its entropy. We realised that in order to answer the titular question of the paper, one must introduce additional structures to the basic measure-theoretic framework known in the literature of scoring rules (Hendrickson and Buehler, 1971). The most important new aspect is the notion of interior and its refinement (known as quasi-interior) in the context of domains with empty interior. Another crucially important idea is to use directional derivatives to describe the subdifferentials of entropy functions. Finally, our approach marks a shift in emphasis from proper scoring rules to a greater focus on entropy functions.

Acknowledgments

Appendix A. Proofs

Lemma A.1 *Let $\mathcal{P}$ be a prediction set and $\Phi : \mathcal{P}^+ \to \mathbb{R}$ be a 1-homogeneous function. If Φ has a (strict) subgradient on $\mathcal{P}^+$, then Φ is a (strictly) sublinear function.*

Proof Let $S : \mathcal{P}^+ \to \operatorname{Lin} \mathcal{P}$ be a (strict) subgradient of Φ. Then S (strictly) satisfies

$$\Phi(p) \geq p \cdot S((1 - \lambda)p + \lambda q)$$
$$\Phi(q) \geq q \cdot S((1 - \lambda)p + \lambda q)$$

for every $p, q \in \mathcal{P}^+$ (p and q not positively collinear), and every $0 < \lambda < 1$. Multiplying the first inequality by $1 - \lambda$, the second one by λ, and then adding them up, we obtain that Φ (strictly) satisfies

$$\Phi(1 - \lambda)p + \lambda q) \leq (1 - \lambda)\Phi(p) + \lambda\Phi(q).$$

$\blacksquare$

Proof [of Lemma 2.1] We first show that $\operatorname{cone}(\mathcal{P}^+ - q) \subset \mathcal{D}(q)$. It is easy to see that $\mathcal{D}(q)$ is closed under taking conic combinations. The claim follows from the fact that $(\mathcal{P}^+ - q) \subset \mathcal{D}(q)$. We now show that $\mathcal{D}(q) \subset \operatorname{cone}(\mathcal{P}^+ - q)$. If $p \in \mathcal{D}(q)$, then there is $\epsilon_p > 0$ and $r \in \mathcal{P}^+$ such that $q + \epsilon_p p = r$. Then $p = (r - q)\epsilon_p^{-1}$ and hence $p \in \operatorname{cone}(\mathcal{P}^+ - q)$. $\blacksquare$

Proof [of Proposition 2.3] (a) For $p \in \mathcal{D}(q)$ arbitrary, consider the line in span $\mathcal{P}$ with parametric equation

$$\gamma(t) = q + t(p - q), \quad t \in \mathbb{R},$$

passing through q and p. Clearly, $\gamma(0) = q$ and $\gamma(1) = p$. Moreover, there is some $\epsilon > 0$ such that the interval $[0, \epsilon]$ is mapped entirely in $\mathcal{P}^+$ under γ (if $p \in \mathcal{P}^+$, then $\epsilon \geq 1$). Then the function

$$\phi(t) = \Phi(q + t(p - q)), \quad t \in [0, \epsilon],$$

is convex and its slope function

$$s_\phi(t_1, t_2) = \frac{\phi(t_2) - \phi(t_1)}{t_2 - t_1}, \quad t_1, t_2 \in [0, \epsilon],$$

is nondecreasing (Rockafellar, 1972; Niculescu and Persson, 2006). We have that

$$\Phi'_+(p, q) = \lim_{t_2 \to 0+} \frac{\phi(t_2) - \phi(0)}{t_2} = \inf_{t_2 > 0} \frac{\phi(t_2) - \phi(0)}{t_2}.$$

If $p \in \mathcal{O}(q)$, then there is some $\delta > 0$ such that the interval $[-\delta, \delta]$ is mapped entirely in $\mathcal{P}^+$ under γ. Let $-\delta \leq t_1 < 0 < t_2 \leq \delta$. To prove that $\Phi'_+(p, q)$ is finite, we consider

$$\frac{\phi(0) - \phi(t_1)}{-t_1} \leq \frac{\phi(t_2) - \phi(0)}{t_2},$$

and take the infimum in t_2.

(b) Homogeneity of $\Phi'_+(\cdot, q)$ follows from:

$$\Phi_+(\lambda p, q) = \lim_{\tau \to 0^+} \frac{\Phi(q + \tau \lambda p) - \Phi(q)}{\tau} \leq \lambda \lim_{\tau \to 0^+} \frac{\Phi(q + \lambda \tau p) - \Phi(q)}{\lambda \tau}$$
$$= \lambda \Phi_+(p, q).$$

Let $p_1, p_2 \in \mathcal{D}(q)$. Subadditivity of $\Phi'_+(\cdot, q)$ follows from:

$$\Phi'_+(p_1 + p_2, q) = \lim_{\tau \to 0^+} \frac{\Phi(q + \tau(p_1 + p_2)) - \Phi(q)}{\tau}$$
$$\leq \lim_{\tau \to 0^+} \frac{\Phi(q/2 + \tau p_1) - \Phi(q)/2}{\tau} + \lim_{\tau \to 0^+} \frac{\Phi(q/2 + \tau p_2) - \Phi(q)/2}{\tau}$$
$$= \lim_{\tau \to 0^+} \frac{\Phi(q + 2\tau p_1) - \Phi(q)}{2\tau} + \lim_{\tau \to 0^+} \frac{\Phi(q + 2\tau p_2) - \Phi(q)}{2\tau}$$
$$= \Phi'_+(p_1, q) + \Phi'_+(p_2, q).$$

(c) The claim follows from

$$\Phi'_+(p, \lambda q) = \lim_{\tau \to 0^+} \frac{\Phi(\lambda q + \tau p) - \Phi(\lambda q)}{\tau} = \lim_{\tau \to 0^+} \frac{\Phi(q + \tau p/\lambda) - \Phi(q)}{\tau/\lambda}$$
$$= \Phi'_+(p, q).$$

(d) We have

$$\Phi(p) \geq \Phi(q+p) - \Phi(q) \geq \frac{\Phi(q+\tau p) - \Phi(q)}{\tau} \geq \Phi'_+(p, q),$$

where $0 < \tau < 1$. The first inequality follows from sublinearity of Φ, while the second and third follow from the fact that the slope function of Φ is nondecreasing. It remains to show that $\Phi(q) = \Phi'_+(q, q)$. This follows immediately from

$$\Phi(q) = \lim_{\tau \to 0^+} \frac{(1+\tau)\Phi(q) - \Phi(q)}{\tau} = \lim_{\tau \to 0^+} \frac{\Phi(q+\tau q) - \Phi(q)}{\tau}$$
$$= \Phi'_+(q, q).$$

(e) The claim is a direct consequence of

$$0 = \Phi'_+(0, q) = \Phi'_+(p - p, q) \leq \Phi'_+(p, q) + \Phi'_+(-p, q).$$

(f) To show that $\mathcal{O}'(q)$ is a linear subspace of $\mathcal{O}(q)$ it is enough to show that it is closed under scalar multiplication and vector addition. Let $\lambda \in \mathbb{R}$ and $p \in \mathcal{O}'(q)$. Then, for $\lambda \geq 0$, $\Phi'_+(\lambda p, q) = \lambda \Phi'_+(p, q)$. Analogously, for $\lambda < 0$ we have

$$\Phi'_+(\lambda p, q) = \Phi'_+(-\lambda(-p), q) = -\lambda \Phi'_+(-p, q) = \lambda(-\Phi'_+(-p, q)) = \lambda \Phi'_+(p, q).$$

Therefore, $\Phi'_+(\lambda p, q) = \lambda \Phi'_+(p, q)$ for any $\lambda \in \mathbb{R}$ and $p \in \mathcal{O}'(q)$. Then multiplying by λ both sides of the identity

$$-\Phi'_+(-p, q) = \Phi'_+(p, q)$$

and using the previous identity, we get that $\lambda p \in \mathcal{O}'(q)$. Hence, $\mathcal{O}'(q)$ is closed under scalar multiplication.

Suppose now that $p, r \in \mathcal{O}'(q)$. We have

$$\Phi'_+(p + r, q) \leq \Phi'_+(p, q) + \Phi'_+(r, q) = -(\Phi'_+(-p, q) + \Phi'_+(-r, q))$$
$$\leq -\Phi'_+(-p - r, q) \leq \Phi'_+(p + r, q),$$

where the last inequality follows from (e). Clearly, we must have equalities throughout. In particular,

$$-\Phi'_+(-p - r, q) = \Phi'_+(p + r, q)$$

and

$$\Phi'_+(p + r, q) = \Phi'_+(p, q) + \Phi'_+(r, q).$$

Hence $p + r \in \mathcal{O}'(q)$. We conclude that $\mathcal{O}'(q)$ is a linear subspace and $\Phi'_+(\cdot, q)\big|_{\mathcal{O}'(q)}$ is linear.
∎

Proof [of Proposition 2.4] (a) The sufficient part of the claim follows from Proposition 2.3 (d). Let us now show the necessary part. To that end, let $q^* \in \operatorname{Lin} \mathcal{P}$ be a subgradient of Φ at q, and let $p \in \mathcal{P}^+$ be arbitrary. Setting $q_t = q + (1-t)p$, we have $\Phi(q_t) \geq q_t \cdot q^*$ for

all $t \in [0, 1]$. Subtracting $\Phi(q)$ from both sides of the inequality and dividing by $(1 - t)$, for $t \in (0, 1)$, we get

$$\frac{\Phi(q + (1 - t)p) - \Phi(q)}{1 - t} \geq p \cdot q^*.$$

Letting $t \uparrow 1$, we get

$$\Phi'_+(p, q) \geq p \cdot q^*$$

as desired.

(b) The claim follows by restricting Φ to 1-dimensional affine spaces through q. On these spaces Φ is convex and differentiable and therefore has a unique subgradient. Since these subspaces cover the whole of span $\mathcal{P}$, it follows that the directional derivative $\Phi'(\cdot, q)$ is the unique subgradient of Φ there.

(c) In view of Proposition 2.3 (a), $\Phi'_+(p, q)$ is finite for each $p \in \mathcal{O}(q) = \text{span} \, \mathcal{P}$. The hypothesis implies that there is at least one 1-dimensional linear subspace of span $\mathcal{P}$ on which $\Phi'_+(\cdot, q)$ is not linear. There are infinitely many ways we can choose a linear function on that space that is dominated by $\Phi'_+(\cdot, q)$. The claim now follows from the Hahn-Banach theorem stated below as Theorem B.4.

(d) Since $\mathcal{O}(q) \neq \text{span} \, \mathcal{P}$, it follows that $\mathcal{P}^+ \setminus \mathcal{O}(q)$ is nonempty. Take any p in that set and consider the 1-dimensional linear space generated by the span of p. Since $\Phi'_+(\cdot, q)$ is defined only on its positive half-space, there are infinitely many linear functions that are dominated by $\Phi'_+(\cdot, q)$ on the whole space. The proof now follows from Theorem B.4.

(e) There is no element of $\text{Lin} \, \mathcal{P}$ that satisfies the condition in part (a) of this proposition. Therefore, $\partial \Phi(q) = \emptyset$. ■

Proof [of Theorem 3.1] Suppose that $q^* \in \mathcal{L}(\mathcal{P})$ satisfies $p \cdot q^* \leq \Phi'_+(p, q)$ for all $p \in \mathcal{P}^+$, with equality for $p = q$. In view of Proposition 2.3 (d), we have that $p \cdot q^* \leq \Phi(p)$ for all $p \in \mathcal{P}^+$, and $q \cdot q^* = \Phi(q)$. Hence, q^* is a $\mathcal{P}$-integrable subgradient of Φ at q.

The converse claim, that is, if q^* is a $\mathcal{P}$-integrable subgradient of Φ at q, then $p \cdot q^* \leq \Phi'_+(p, q)$ for all $p \in \mathcal{P}^+$, with equality for $p = q$, follows from Proposition 2.4 (a). ■

Proof [of Theorem 3.2] The hypothesis implies that $\Phi'_+(\cdot, q)$ is linear on $\mathcal{O}(q) \subset \mathcal{P}^+$. By restricting Φ to 1-dimensional subspaces of $\mathcal{O}(q)$ it follows immediately that any subgradient of Φ must agree with q^* on $\mathcal{O}(q)$. The assumption that $\mathcal{O}(q)^\perp = \{0\}$ implies that Φ may have at most one $\mathcal{P}$-integrable subgradient at q. Then the claim follows from the fact that q^* is a subgradient of Φ at q. ■

Appendix B. Some Additional Facts

The positive cones in many standard function spaces are nowhere dense sets. Let us show this for the Lebesgue space $L^1(\mathbb{R}^d)$. The positive cone of $L^1(\mathbb{R}^d)$ consists of all Lebesgue integrable functions $f \geq 0$ a.e. on $\mathbb{R}^d$ and is denoted by $L^1_+(\mathbb{R}^d)$. We recall that a set in a topological vector space is *nowhere dense* if its closure has empty interior.

Proposition B.1 *The positive cone of $L^1(\mathbb{R}^d)$ is nowhere dense.*

Proof We show that for every $f \geq 0$ a.e., there is $g \geq 0$ a.e. such that, for every $\alpha > 0$, $f - \alpha g \notin L^1_+(\Omega)$. This means that no open ball about f is contained in $L^1_+(\mathbb{R}^d)$. Since $L^1_+(\mathbb{R}^d)$ is closed, then this would imply that $L^1_+(\mathbb{R}^d)$ is nowhere dense.

To prove our claim, we use the fact that there is no absolutely convergent series with a slowest rate of decay at infinity. We begin by partitioning $\mathbb{R}^d$ into dyadic regions

$$\omega_k = \{2^k \leq |x| < 2^{k+1}\}$$

for $k \in \mathbb{Z}$. For $f \in L^1(\mathbb{R}^d)$, we set

$$a_k = \int_{\omega_k} f(x)dx.$$

We have that the series

$$\sum_{k=0}^{\infty} a_k = \int_{\mathbb{R}^d} f(x)dx$$

is absolutely convergent. If $r_k = \sum_{i \geq k} a_i$ is the tail of the series for each k, then the series $\sum_{k \geq 0} a_k/\sqrt{r_k}$ is also convergent (Rudin, 1976). Notice that the ratio of the common term of the second to the first series tends to infinity as $k \to \infty$. Therefore, the second series has a strictly slower rate of convergence. There exists a function $g \in L^1_+(\mathbb{R}^d)$ such that the integrals of g on ω_k are $b_k = a_k/\sqrt{r_k}$ and

$$\sum_{k=0}^{\infty} b_k = \int_{\mathbb{R}^d} g(x)dx.$$

Clearly, for any $\alpha > 0$, the difference $f - \alpha g$ changes sign for some ω_k, and hence $f - \alpha g \notin L^1_+(\mathbb{R}^d)$. ∎

The next example illustrates the notion of topological subgradient in the case when the assumption $N^* \subset \mathcal{L}(\mathcal{P})$ is not satisfied.

Example B.2 Consider a Lebesgue measure space $(\Omega, \mathcal{A}, \mu)$ with Ω a compact subset of $\mathbb{R}^d$. We set $\mathcal{P}^+$ to be the positive cone of $C(\Omega)$, that is, the set of all nonnegative continuous functions on Ω. The continuous dual of $C(\Omega)$ is the space of all real-valued Radon measures on Ω. The fact that $\mathcal{P}^+$ contains constants implies that $\mathcal{L}(\mathcal{P}) \subseteq L^1(\Omega)$. Actually, $\mathcal{L}(\mathcal{P}) = L^1(\Omega)$ and hence the $\mathcal{P}$-integrable functions are the Radon measures that have a Lebesgue density. Since $L^1(\Omega) \subsetneq (C(\Omega))^*$, we see that in this case the notion of a $\mathcal{P}$-integrable subgradient is more restrictive than that of a topological subgradient.

We proceed to examine the implications of the latter observation on a concrete sublinear function. Let $\Phi : C(\Omega) \to \mathbb{R}$ be the *supremum function*, that is,

$$\Phi(p) = \sup_{x \in \Omega} p(x).$$

It is easy to check that Φ is non-strictly sublinear and continuous. The supporting hyperplane theorem guarantees the existence of a topological subgradient of Φ at each point in its

domain that is a *real Radon measure.* Let us see whether the subgradient is regular enough to be a proper scoring rule.

We first demonstrate that there are points $q \in \mathcal{P}^+$ at which Φ has no subgradient in $\mathcal{L}(\mathcal{P})$. To that end, let $\mathcal{M}(q)$ denote the set of *modes* of q, that is, the subset of Ω where q reaches its maximum. Notice that $\mathcal{M}(q)$ is always compact. It can be shown that

$$\Phi'_+(p, q) = \sup_{x \in \mathcal{M}(q)} p(x),$$

the proof of which is left to the reader. When $\mathcal{M}(q) = \{x_0\}$ is a singleton, $\Phi'_+(\cdot, q) = \delta(x - x_0)$ is Dirac's delta function. Clearly, in this case Φ is Gâteaux differentiable with derivative $\delta(x - x_0)$. We claim that Φ has no $\mathcal{P}$-integrable subgradient for any density q with $\mu(\mathcal{M}(q)) = 0$.

Suppose conversely that $q^* \in \mathcal{L}(\mathcal{P})$, $q^* \neq 0$, is a subgradient of Φ at q. Then

$$\Phi'_+(p, q) \geq p \cdot q^*$$

for all $p \in \mathcal{P}^+$. We shall show that this inequality implies $q^*(x) \leq 0$ a.e. on Ω, which leads to a contradiction with $\Phi(q) = q \cdot q^* > 0$.

To show the latter claim, notice that $\Omega \setminus \mathcal{M}(q)$ is open, and hence for any $y \in \Omega \setminus \mathcal{M}(q)$, there is $\epsilon_y > 0$ such that the ball about y of radius ϵ_y lies in the complement of $\mathcal{M}(q)$ with respect to Ω. Let $\{p_k\}$ be a sequence of densities approximating $\delta(x - y)$ entirely supported on this ball. Since $\Phi'_+(p_k, q) = 0$, we get that $p_k \cdot q^* \leq 0$. If y is a Lebesgue point of q^*, then we have the limit

$$\lim_{k \to \infty} p_k \cdot q^* = \delta(\cdot - y) \cdot q^* = q^*(y).$$

Since almost every point of q^* is a Lebesgue point, we get that $q^*(x) \leq 0$ a.e. on Ω. This completes the proof of the claim.

In the case $\mu(\mathcal{M}(q)) > 0$, we may find a $\mathcal{P}$-integrable subgradient of Φ at q. Consider the function

$$q^*(x) = \begin{cases} \frac{1}{\mu(\mathcal{M}(q))} & x \in \mathcal{M}(q) \\ 0 & x \in \Omega \setminus \mathcal{M}(q). \end{cases}$$

Clearly, $q \cdot q^* = \sup_{x \in \Omega} q(x)$ and $p \cdot q^* \leq \sup_{x \in \Omega} p(x)$ for all $p \in \mathcal{P}^+$. This furnishes our claim.

In our final example, we illustrate the fact that at boundary points a sublinear function has either no subgradient, or infinitely many.

Example B.3 Take $\Phi(x, y) = x + y$ on $\mathbb{R}^2_+ = \{(x, y) \mid x \geq 0, y \geq 0\}$. The graph of Φ is a part of a plane, so it is easy to see that Φ has infinitely many supporting planes at the boundaries of $\mathbb{R}^2_+$. Consider now

$$\Phi(x, y) = x \ln \frac{x}{x + y} + y \ln \frac{y}{x + y}$$

on $\mathbb{R}^2_+$, which is Shannon entropy for binary variables. A computation shows that

$$\nabla \Phi(x, y) = \ln \frac{x}{x + y} + \ln \frac{y}{x + y}$$

and hence $\nabla\Phi(x, y) \to -\infty$ when (x, y) tends to the boundary of $\mathbb{R}_+^2$. This means that Φ has vertical tangent planes through the coordinate axes, which implies that Φ has no subgradient on the boundary of its domain.

The situation is the same when $\mathcal{P}^+$ is a subset of an infinite dimensional vector space. For example, one may use the Hahn-Banach theorem presented below to show the existence of multiple supporting hyperplanes at boundary points q for which $\Phi'_+(p, q)$ is finite for all $p \in \mathcal{P}^+$. If, instead, there is $p \in \mathcal{P}^+$ for which $\Phi'_+(p, q) = -\infty$, then Φ has no subgradient at q.

We now state a slight generalisation of the classical Hahn-Banach theorem. Let E be a real vector space and $K \subset E$ be a convex cone.

Theorem B.4 (Hahn-Banach theorem) *Let $\phi : K \to \mathbb{R}$ be a sublinear function and $l_0 : E_0 \to \mathbb{R}$ be a linear functional on a linear subspace $E_0 \subseteq E$ which is dominated by ϕ on $E_0 \cap K$, i.e.*

$$l_0(q) \leq \phi(q), \qquad \forall q \in E_0 \cap K.$$

Then there exists a linear extension $l : E \to \mathbb{R}$ of l_0 to the whole space E such that

$$l(q) = l_0(q), \qquad \forall q \in E_0,$$
$$l(q) \leq \phi(q), \qquad \forall q \in E \cap K.$$

In the classical formulation of the theorem, we have $K = E$. The proof of the version with $K \subset E$ is the same. In fact, if anything, the condition $K \subset E$ is easier to satisfy than $K = E$ when extending l_0.

References

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

J. Borwein and J. Vanderwerff. *Convex Functions: Constructions, Characterizations and Counterexamples*. Number 109 in Encyclopedia of Mathematics and its Applications. Cambridge University Press, Cambridge, 2010.

J. M. Borwein and A. S. Lewis. Partially finite convex programming, Part I: Quasi relative interiors and duality theory. *Mathematical Programming*, 57(1–3):15–48, 1992.

G. W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

A. P. Dawid. The geometry of proper scoring rules. *Annals of the Institute of Statistical Mathematics*, 59:77–93, 2007.

A. P. Dawid and M. Musio. Estimation of spatial processes using local scoring rules. *AStA Advances in Statistical Analysis*, 96:1–7, 2012. Spatial special issue.

A. P. Dawid and M. Musio. Theory and applications of proper scoring rules. *Metron*, 72: 169–183, 2014.

A. P. Dawid, S. Lauritzen, and M. Parry. Proper local scoring rules on discrete sample spaces. *The Annals of Statistics*, 40(1):593–608, 2012.

W. Ehm and T. Gneiting. Local proper scoring rules of order two. *The Annals of Statistics*, 40(1):609–637, 2012.

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

R. M. Frongillo and I. Kash. General truthfulness characterizations via convex analysis. In *Lecture Notes in Computer Science*, volume 8877 of *Web and Internet Economics*, pages 354–370. Springer, 2014.

R. E. Fullerton and C. C. Braunschweiger. Quasi-interior points of cones. *Technical Report 2, University of Delaware, Newark, Delaware*, 1963.

T. Gneiting and M. Katzfuss. Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1:125–151, 2014.

T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102:359–376, 2007.

P. D. Grünwald and A. P. Dawid. Game theory, maximum entropy, minimum discrepancy, and robust Bayesian decision theory. *The Annals of Statistics*, 32(4):1367–1433, 2004.

A. Hendrickson and R. Buehler. Proper scores for probability forecasters. *The Annals of Mathematical Statistics*, 42(6):1916–1921, 1971.

A. Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6:695–709, 2005.

A. Hyvärinen. Some extensions of score matching. *Computational Statistics & Data Analysis*, 51:2499–2512, 2007.

T. Kanamori and H. Fujisawa. Affine invariant divergences associated with proper composite scoring rules and their applications. *Bernoulli*, 20(4):2278–2304, 2014.

T. Kanamori and H. Fujisawa. Robust estimation under heavy contamination using enlarged models. *Biometrika*, 2015. doi:10.1093/biomet/asv014.

C. Niculescu and L.-E. Persson. *Convex Functions and Their Applications*. CMS Books in Mathematics. Springer, 2006.

M. Parry, A. P. Dawid, and S. Lauritzen. Proper local scoring rules. *The Annals of Statistics*, 40(1):561–592, 2012.

R. T. Rockafellar. *Convex Analysis.* Princeton Mathematical Series. Princeton University Press, second edition, 1972.

W. Rudin. *Functional Analysis.* McGraw-Hill, 1973.

W. Rudin. *Principles of Mathematical Analysis.* International Series in Pure and Applied Mathematics. McGraw-Hill, third edition, 1976.

P. Sánchez-Moreno, A. Zarzo, and J. S. Dehesa. Jensen divergence based on Fisher's information. *Journal of Physics A: Mathematical and Theoretical*, 45, 2012. 125305.

R. C. Williamson. The geometry of losses. *Conference On Learning Theory (COLT)*, 35: 1078–1108, 2014.

C. Zalinescu. *Convex Analysis in General Vector Spaces.* World Scientific, 2002.