# Probabilistic coherence and proper scoring rules[*]

Joel Predd
Rand Corporation

Robert Seiringer
Princeton University

Elliott H. Lieb
Princeton University

Daniel Osherson
Princeton University

H. Vincent Poor
Princeton University

Sanjeev Kulkarni
Princeton University

April 15, 2009

**Abstract**

We provide self-contained proof of a theorem relating probabilistic coherence of forecasts to their non-domination by rival forecasts with respect to any proper scoring rule. The theorem recapitulates insights achieved by other investigators, and clarifies the connection of coherence and proper scoring rules to Bregman divergence.

## 1 Introduction

Scoring rules measure the quality of a probability-estimate for a given event, with lower scores signifying probabilities that are closer to the event's status (1 if it occurs, 0 otherwise). The sum of the scores for estimates $\mathbf{p}$ of a vector $\mathcal{E}$ of events is called the "penalty" for $\mathbf{p}$. Consider two potential defects in $\mathbf{p}$.

- There may be rival estimates $\mathbf{q}$ for $\mathcal{E}$ whose penalty is guaranteed to be lower than the one for $\mathbf{p}$, regardless of which events come to pass.

[*] © 2009 by the authors. This paper may be reproduced, in its entirety, for non-commercial purposes. Research supported by NSF grants PHY-0652854 to Lieb, and PHY-0652356 to Seiringer. R.S. also acknowledges partial support from an A. P. Sloan Fellowship. D.O. thanks the Henry Luce Foundation for support. H.V.P. acknowledges support from NSF grants ANI-03-38807 and CNS-06-25637. S.K. aknowledges support from the Office of Naval Research under contract numbers W911NF-07-1-0185 and N00014-07-1-0555.

- The events in $\mathcal{E}$ may be related by inclusion or partition, and $\mathbf{p}$ might violate constraints imposed by the probability calculus (for example, that the estimate for an event not exceed the estimate for any event that includes it).

Building on the work of earlier investigators (see below), we show that for a broad class of scoring rules known as "proper" the two defects are equivalent. An exact statement appears as Theorem 1. To reach it, we first explain key concepts intuitively (the next section) then formally (Section 3). Proof of the the theorem proceeds via three propositions of independent interest (Section 4). We conclude with generalizations of our results and an open question.

## 2 Intuitive account of concepts

Imagine that you attribute probabilities .6 and .9 to events $\mathsf{E}$ and $\mathsf{F}$, respectively, where $\mathsf{E} \subseteq \mathsf{F}$. It subsequently turns out that $\mathsf{F}$ comes to pass but not $\mathsf{E}$. How shall we assess the perspicacity of your two estimates, which may jointly be called a **probabilistic forecast**? According to one method (due to Brier, 1950) truth and falsity are coded by 1 and 0, and your estimate of the chance of $\mathsf{E}$ is assigned a score of $(0 - .6)^2$ since $\mathsf{E}$ did not come true (so your estimate should ideally have been zero). Your estimate for $\mathsf{F}$ is likewise assigned $(1 - .9)^2$ since it should have been one. The sum of these numbers serves as overall **penalty**.

Let us calculate your expected penalty for $\mathsf{E}$ (prior to discovering the facts). With .6 probability you expected a score of $(1 - .6)^2$, and with the remaining probability you expected a score of $(0 - .6)^2$, hence your overall expectation was $.6(1 - .6)^2 + .4(0 - .6)^2 = .24$. Now suppose that you attempted to improve (lower) this expectation by insincerely announcing .65 as the chance of $\mathsf{E}$, even though your real estimate is .6. Then your expected penalty would be $.6(1 - .65)^2 + .4(0 - .65)^2 = .2425$, worse than before. Differential calculus reveals the general fact:

> Suppose your probability for an event $\mathsf{E}$ is $\mathsf{p}$, that your announced probability is $\mathsf{x}$, and that your penalty is assessed according to the rule: $(1-\mathsf{x})^2$ if $\mathsf{E}$ comes out true; $(0-\mathsf{x})^2$ otherwise. Then your expected penalty is uniquely minimized by choosing $\mathsf{x} = \mathsf{p}$.

Our scoring rule thus encourages sincerity since your interest lies in announcing probabilities that conform to your beliefs. Rules like this are called **strictly proper**.[1] (We

[1] For brevity, the term "proper" will be employed instead of the usual "strictly proper".

add a continuity condition in our formal treatment, below.) For an example of an improper rule, substitute absolute deviation for squared deviation in the original scheme. According to the new rule, your expected penalty for $\mathsf{E}$ is $.6|1 - .6| + .4|0 - .6| = .48$ whereas it drops to $.6|1 - .65| + .4|0 - .65| = .47$ if you fib as before.

Consider next the rival forecast of .95 for $\mathsf{E}$ and .55 for $\mathsf{F}$. Because $\mathsf{E} \subseteq \mathsf{F}$, this forecast is inconsistent with the probability calculus (or **incoherent**). Table 1 shows that the original forecast **dominates** the rival inasmuch as its penalty is lower however the facts play out. This association of incoherence and domination is not an accident. No matter what proper scoring rule is in force, any incoherent forecast can be replaced by a coherent one whose penalty is lower in every possible circumstance; there is no such replacement for a coherent forecast. This fact is formulated as Theorem 1 in the next section. It can be seen as partial vindication of probability as an expression of chance.[2]

| | Forecast | |
|---|---|---|
| Logical possibilities when $\mathsf{E} \subseteq \mathsf{F}$ | original | rival |
| E true<br>F true | .17 | .205 |
| E false<br>F true | .37 | 1.105 |
| E false<br>F false | 1.17 | 1.205 |

Table 1: Penalties for two forecasts in alternative possible realities

These ideas have been discussed before, first by de Finetti (1974) who began the investigation of dominated forecasts and probabilistic consistency (called **coherence**). His work relied on the **quadratic scoring rule**, introduced above.[3] Lindley (1982) generalized de Finetti's theorem to a broad class of scoring rules. Specifically, he proved that for every sufficiently regular generalization $\mathsf{s}$ of the quadratic score, there is a transformation $\mathsf{T} : \mathfrak{R} \to \mathfrak{R}$ such that a forecast $\mathbf{f}$ is not dominated by any other forecast with respect to $\mathsf{s}$ if and only if the transformation of $\mathbf{f}$ by $\mathsf{T}$ is probabilistically coherent. It has been suggested to us that Theorem 1 below can be perceived in Lindley's discussion, especially in his Comment 2 (p. 7), which deals with scoring rules that he qualifies as proper. We are agreeable to crediting Lindley with the theorem (under somewhat different regularity conditions) but it seems to us that his discussion is clouded by reliance on the transformation $\mathsf{T}$ to define proper scoring rules and to state the main result.

In any event, fresh insight into proper scoring rules comes from relating them to a generalization of metric distance known as **Bregman divergence** (Bregman, 1967). This relationship was studied by Savage (1971), albeit implicitly, and more recently by Banerjee et al. (2005) and Gneiting and Raftery (2007). So far as we know, their results

[2] The other classic vindication involves sure-loss contracts; see Skyrms (2000).

[3] For analysis of de Finetti's work, see Joyce, 1998. Note that some authors use the term **inadmissible** to qualify dominated forecasts.

have yet to be connected to the issue of dominance. The connection is explored here.

More generally, to pull together the threads of earlier discussions, the present work offers a self-contained account of the relations among (i) coherent forecasts, (ii) Bregman divergences, and (iii) domination with respect to proper scoring rules. Only elementary analysis is presupposed. We begin by formalizing the concepts introduced above.[4]

## 3 Framework and Main Result

Let $\Omega$ be a nonempty **sample space**. Subsets of $\Omega$ are called **events**. Let $\mathcal{E}$ be a vector $(E_1, \cdots, E_n)$ of $n \geqslant 1$ events over $\Omega$. We assume that $\Omega$ and $\mathcal{E}$ have been chosen and are now fixed for the remainder of the discussion. We require $\mathcal{E}$ to have finite dimension $n$ but otherwise our results hold for any choice of sample space and events. In particular, $\Omega$ can be infinite. We rely on the usual notation $[0,1]$, $(0,1)$, $\{0,1\}$ to denote, respectively, the closed interval $\{x : 0 \leqslant x \leqslant 1\}$, the open interval $\{x : 0 < x < 1\}$ and the two-point set containing $0, 1$.

**Definition 1.** Any element of $[0,1]^n$ is called a **(probability) forecast (for $\mathcal{E}$)**. A forecast $\mathbf{f}$ is **coherent** just in case there is a probability measure $\mu$ over $\Omega$ such that for all $i \leqslant n$, $f_i = \mu(E_i)$.

A forecast is thus a list of $n$ numbers drawn from the unit interval. They are interpreted as claims about the chances of the corresponding events in $\mathcal{E}$. The first event in $\mathcal{E}$ is assigned the probability given by the first number ($f_1$) in $\mathbf{f}$, and so forth. A forecast is coherent if it is consistent with some probability measure over $\Omega$.

This brings us to scoring rules. In what follows, the numbers 0 and 1 are used to represent falsity and truth, respectively.

**Definition 2.** A function $s : \{0,1\} \times [0,1] \to [0,\infty]$ is said to be a **proper scoring rule** in case

(a) $ps(1,x) + (1-p)s(0,x)$ is uniquely minimized at $x = p$ for all $p \in [0,1]$.

(b) $s$ is continuous, meaning that for $i \in \{0,1\}$, $\lim_{n\to\infty} s(i, x_n) = s(i,x)$ for any sequence $x_n \in [0,1]$ converging to $x$.

For condition 2(a), think of $p$ as the probability you have in mind, and $x$ as the one you announce. Then $ps(1,x) + (1-p)s(0,x)$ is your expected score. Fixing $p$ (your genuine

[4] For application of scoring rules to the assessment of opinion, see Gneiting and Raftery (2007) along with Bernardo and Smith (1994, §2.7.2) and references cited there.

belief), the latter expression is a function of the announcement $x$. Proper scoring rules encourage candor by minimizing the expected score exactly when you announce $p$.

The continuity condition is consistent with $s$ assuming the value $+\infty$. This can only occur for the arguments $(0,1)$ or $(1,0)$, representing categorically mistaken judgment. For if $s(0,p)=\infty$ for some $p \neq 1$, then $ps(1,x)+(1-p)s(0,x)$ can not have a unique minimum at $x=p$; similarly, $s(1,p) < +\infty$ for $p \neq 0$. An interesting example of an unbounded proper scoring rule (Good, 1952) is

$$s(i,x) = -\ln|1-i-x| \ .$$

A comparison of alternative rules is offered in Selten (1998).

For an event $E$, we let $C_E$ be the characteristic function of $E$; that is, for all $\omega \in \Omega$, $C_E(\omega)=1$ if $\omega \in E$ and 0 otherwise. Intuitively, $C_E(\omega)$ reports whether $E$ is true or false if Nature chooses $\omega$.

**Definition 3.** Given proper scoring rule $s$, the **penalty** $P_s$ based on $s$ for forecast $\mathbf{f}$ and $\omega \in \Omega$ is given by:

$$P_s(\omega,\mathbf{f}) = \sum_{i \leqslant n} s(C_{E_i}(\omega), f_i). \qquad (1)$$

Thus, $P_s$ sums the scores (conceived as penalties) for all the events under consideration. Henceforth, the proper scoring rule $s$ is regarded as given and fixed. The theorem below holds for any choice we make.

**Definition 4.** Let a forecast $\mathbf{f}$ be given.

(a) $\mathbf{f}$ is **weakly dominated** by a forecast $\mathbf{g}$ in case $P_s(\omega,\mathbf{g}) \leqslant P_s(\omega,\mathbf{f})$ for all $\omega \in \Omega$.

(b) $\mathbf{f}$ is **strongly dominated** by a forecast $\mathbf{g}$ in case $P_s(\omega,\mathbf{g}) < P_s(\omega,\mathbf{f})$ for all $\omega \in \Omega$.

Strong domination by a rival, coherent forecast $\mathbf{g}$ is the price to be paid for an incoherent forecast $\mathbf{f}$. Indeed, we shall prove the following version of Lindley (1982), Comment 2.

**Theorem 1.** Let a forecast $\mathbf{f}$ be given.

(a) If $\mathbf{f}$ is coherent then it is not weakly dominated by any forecast $\mathbf{g} \neq \mathbf{f}$.

(b) If $\mathbf{f}$ is incoherent then it is strongly dominated by some *coherent* forecast $\mathbf{g}$.

Thus, if $\mathbf{f}$ and $\mathbf{g}$ are coherent and $\mathbf{f} \neq \mathbf{g}$ then neither weakly dominates the other. The theorem follows from three propositions of independent interest, stated in the next section. We close the present section with a corollary.

**Corollary 1.** A forecast $\mathsf{f}$ is weakly dominated by a forecast $\mathsf{g} \neq \mathsf{f}$ if and only if $\mathsf{f}$ is strongly dominated by a coherent forecast.

*Proof of Corollary 1.* The right-to-left direction is immediate from Definition 4. For the left-to-right direction, suppose forecast $\mathsf{f}$ is weakly dominated by some $\mathsf{g} \neq \mathsf{f}$. Then by Theorem 1(a), $\mathsf{f}$ is not coherent. So by Theorem 1(b), $\mathsf{f}$ is strongly dominated by some coherent forecast. □

## 4 Three Propositions

The first proposition is a characterization of coherence. It is due to de Finetti (1974).

**Definition 5.** Let $V = \{(C_{E_1}(\omega), \cdots, C_{E_n}(\omega)) : \omega \in \Omega\} \subseteq \{0,1\}^n$. Let the cardinality of $V$ be $k$. Let *conv*$(V)$ be the **convex hull** of $V$, i.e., *conv*$(V)$ consists of all vectors of form $a_1\boldsymbol{v}_1 + \cdots + a_k\boldsymbol{v}_k$, where $\boldsymbol{v}_i \in V$, $a_i \geqslant 0$, and $\sum_{i=1}^{k} a_i = 1$.

The $E_i$ may be related in various ways, so $k < 2^n$ is possible (indeed, this is the case of interest).

**Proposition 1.** A forecast $\mathsf{f}$ is coherent if and only if $\mathsf{f} \in$ *conv*$(V)$.

The next proposition characterizes scoring rules in terms of convex functions. Recall that a convex function $\varphi$ on a convex subset of $\mathfrak{R}^n$ satisfies $\varphi(a\boldsymbol{x} + (1-a)\boldsymbol{y}) \leqslant a\varphi(\boldsymbol{x}) + (1-a)\varphi(\boldsymbol{y})$ for all $0 < a < 1$ and all $\boldsymbol{x}$, $\boldsymbol{y}$ in the subset. *Strict* convexity means that the inequality is strict unless $\boldsymbol{x} = \boldsymbol{y}$. Variants of the following fact are proved in Savage (1971), Banerjee et al. (2005), and Gneiting and Raftery (2007).

**Proposition 2.** Let $s$ be a proper scoring rule. Then the function $\varphi : [0,1] \to \mathfrak{R}$ defined by $\varphi(x) = -xs(1,x) - (1-x)s(0,x)$ is a bounded, continuous and strictly convex function, differentiable for $x \in (0,1)$. Moreover,

$$s(i,x) = -\varphi(x) - \varphi'(x)(i-x) \quad \forall x \in (0,1)\,. \tag{2}$$

Conversely, if a function $s$ satisfies (2), with $\varphi$ bounded, strictly convex and differentiable on $(0,1)$, and $s$ is continuous on $[0,1]$, then $s$ is a proper scoring rule.

We note that the right side of (2), which is only defined for $x \in (0,1)$, can be continuously extended to $x = 0, 1$. This is the content of the Lemma 1 in the next section. If the extended $s$ satisfies (2) then:

$$s(0,0) = -\varphi(0) \quad \text{and} \quad s(1,1) = -\varphi(1)\,. \tag{3}$$

Finally, our third proposition concerns a well known property of Bregman divergences (see, e.g., Censor and Zenios, 1997). When we apply the proposition to the proof of Theorem 1, $\mathsf{C}$ will be the unit cube in $\mathfrak{R}^n$.

**Definition 6.** Let $\mathsf{C}$ be a convex subset of $\mathfrak{R}^n$ with non-empty interior. Let $\Phi : \mathsf{C} \to \mathfrak{R}$ be a strictly convex function, differentiable in the interior of $\mathsf{C}$, whose gradient $\nabla\Phi$ extends to a bounded, continuous function on $\mathsf{C}$. For $\mathbf{x}, \mathbf{y} \in \mathsf{C}$, the **Bregman divergence** $d_\Phi : \mathsf{C} \times \mathsf{C} \to \mathfrak{R}$ corresponding to $\Phi$ is given by

$$d_\Phi(\mathbf{y}, \mathbf{x}) = \Phi(\mathbf{y}) - \Phi(\mathbf{x}) - \nabla\Phi(\mathbf{x}) \cdot (\mathbf{y} - \mathbf{x}).$$

Because of the strict convexity of $\Phi$, $d_\Phi(\mathbf{y}, \mathbf{x}) \geqslant 0$ with equality if and only if $\mathbf{y} = \mathbf{x}$.

**Proposition 3.** Let $d_\Phi : \mathsf{C} \times \mathsf{C} \to \mathfrak{R}$ be a Bregman divergence, and let $\mathsf{Z} \subseteq \mathsf{C}$ be a closed convex subset of $\mathfrak{R}^n$. For $\mathbf{x} \in \mathsf{C} \setminus \mathsf{Z}$, there exists a unique $\boldsymbol{\pi}_\mathbf{x} \in \mathsf{Z}$, called the **projection of $\mathbf{x}$ onto** $\mathsf{Z}$, such that

$$d_\Phi(\boldsymbol{\pi}_\mathbf{x}, \mathbf{x}) \leqslant d_\Phi(\mathbf{y}, \mathbf{x}) \quad \forall \mathbf{y} \in \mathsf{Z}\,.$$

Moreover,

$$d_\Phi(\mathbf{y}, \boldsymbol{\pi}_\mathbf{x}) \leqslant d_\Phi(\mathbf{y}, \mathbf{x}) - d_\Phi(\boldsymbol{\pi}_\mathbf{x}, \mathbf{x}) \quad \forall \mathbf{y} \in \mathsf{Z}, \mathbf{x} \in \mathsf{C} \setminus \mathsf{Z}\,. \tag{4}$$

Its worth observing that Proposition 3 also holds if $\mathbf{x} \in \mathsf{Z}$, in which case $\boldsymbol{\pi}_\mathbf{x} = \mathbf{x}$ and (4) is trivially satisfied.

## 5 Proofs of Propositions 1–3

*Proof of Proposition 1.* Recall that $n$ is the dimension of $\mathcal{E}$, and that $k$ is the number of elements in $\mathsf{V}$. Let $\mathsf{X}$ be the collection of all nonempty sets of form $\bigcap_{i=1}^n \mathsf{E}_i^*$, where $\mathsf{E}_i^*$ is either $\mathsf{E}_i$ or its complement. ($\mathsf{X}$ corresponds to the minimal non-empty regions appearing in the Venn diagram of $\mathcal{E}$.) It is easy to see that:

(a) $\mathsf{X}$ partitions $\Omega$.

It is also clear that there is a one-to-one correspondence between $\mathsf{X}$ and $\mathsf{V}$ with the property that $e \in \mathsf{X}$ is mapped to $\mathbf{v} \in \mathsf{V}$ such that for all $i \leqslant n$, $e \subseteq \mathsf{E}_i$ iff $v_i = 1$. (Here, $v_i$ denotes the $i$th component of $\mathbf{v}$.) Thus, there are $k$ elements in $\mathsf{X}$. We enumerate them as $e_1, \cdots, e_k$, and the corresponding $\mathbf{v}$ by $\mathbf{v}(e_j)$. Plainly, for all $i \leqslant n$, $\mathsf{E}_i$ is the disjoint union of $\{e_j : j \leqslant k \ \wedge \ v(e_j)_i = 1\}$, and hence:

(b) For any measure $\mu$ , $\mu(E_i) = \sum_{j=1}^{k} \mu(e_j)\nu(e_j)_i$ for all $1 \leqslant i \leqslant n$.

For the left-to-right direction of the proposition, suppose that forecast $f$ is coherent via probability measure $\mu$. Then $f_i = \mu(E_i)$ for all $i \leqslant n$ and hence by (b), $f_i = \sum_{j=1}^{k} \mu(e_j)\nu(e_j)_i$. But the $\mu(e_j)$ are non-negative and sum to one by (a), which shows that $\mathbf{f} \in \mathit{conv}(V)$.

For the converse, suppose that $\mathbf{f} \in \mathit{conv}(V)$, which means that there are non-negative $a_j$'s, with $\sum_j a_j = 1$, such that $\mathbf{f} = \sum_{j=1}^{k} a_j \boldsymbol{\nu}(e_j)$. Let $\mu$ be some probability measure such that $\mu(e_j) = a_j$ for all $j \leqslant k$. By (a) and the assumption about the $a_i$, it is clear that such a measure $\mu$ exists. For all $i \leqslant n$, $f_i = \sum_{j=1}^{k} a_j\nu(e_j)_i = \sum_{j=1}^{k} \mu(e_j)\nu(e_j)_i = \mu(E_i)$ by (b), thereby exhibiting $\mathbf{f}$ as coherent. □

Before giving the proof of Proposition 2, we state and prove the following technical Lemma.

**Lemma 1.** Let $\varphi : [0,1] \to \mathfrak{R}$ be bounded, convex and differentiable on $(0,1)$. Then the limits $\lim_{p\to 0,1} \varphi(p)$ and $\lim_{p\to 0,1} \varphi'(p)$ exist, the latter possibly being equal to $-\infty$ at $x = 0$ or $+\infty$ at $x = 1$. Moreover,

$$\lim_{p\to 0} p\varphi'(p) = \lim_{p\to 1} \varphi'(p)(1-p) = 0\,. \tag{5}$$

*Proof of Lemma 1.* Since $\varphi$ is convex, the limits $\lim_{p\to 0,1} \varphi(p)$ exist, and they are finite since $\varphi$ is bounded. Moreover, $\varphi'$ is a monotone increasing function, and hence also $\lim_{p\to 0,1} \varphi'(p)$ exists (but possibly equals $-\infty$ at $x = 0$ or $+\infty$ at $x = 1$). Finally, Eq. (5) follows again from monotonicity of $\varphi'$ and boundedness of $\varphi$, using that $0 = \lim_{p\to 0} \int_0^p \varphi'(q)dq \leqslant \lim_{p\to 0} p\varphi'(p)$, and likewise at $p = 1$. □

*Proof of Proposition 2.* Let $s$ be a proper scoring rule. For $0 < p < 1$, let

$$\varphi(p) = -\min_x \{ps(1,x) + (1-p)s(0,x)\}\,. \tag{6}$$

By Definition 2(a), the minimum in (6) is achieved at $x = p$, hence $\varphi(p) = -ps(1,p) - (1-p)s(0,p)$.

As a minimum over linear functions, $-\varphi$ is concave; hence $\varphi$ is convex. Clearly, $\varphi$ is bounded (because $s \geqslant 0$ implies, from (6), that $\varphi \leqslant 0$, but a convex function can become unbounded only by going to $+\infty$).

The fact that the minimum is achieved uniquely (Def. 2) implies that $\varphi$ is strictly convex for the following reason. We take $x$, $y \in (0,1)$ and $0 < a < 1$ and set $z =$

$ax + (1-a)y$. Then $\varphi(y) = -y\, s(1,y) - (1-y)\, s(0,y) > -y\, s(1,x) - (1-y)\, s(0,z)$ by uniqueness of the minimizer at $y \neq z$. Similarly, $\varphi(x) = -x\, s(1,x) - (1-x)\, s(0,x) > -x\, s(1,z) - (1-x)\, s(0,z)$. By adding $a$ times the first inequality to $1-a$ times the second we obtain $a\varphi(y) + (1-a)\varphi(x) > -z\, s(1,z) - (1-z)\, s(0,z) = \varphi(z)$, which is precisely the statement of strict convexity.

Let $\psi(p) = s(0,p) - s(1,p)$. If $\varphi$ is differentiable and $\varphi'(p) = \psi(p)$ for all $0 < p < 1$, then (2) is satisfied, as simple algebra shows.

We shall now show that $\varphi$ is, in fact, differentiable and $\varphi' = \psi$. For any $p \in (0,1)$ and small enough $\epsilon$, we have

$$\frac{1}{\epsilon}\left(\varphi(p+\epsilon) - \varphi(p)\right) = \psi(p) \\ - \frac{1}{\epsilon}\left[(p+\epsilon)\left(s(1,p+\epsilon) - s(1,p)\right) + (1-p-\epsilon)\left(s(0,p+\epsilon) - s(0,p)\right)\right] .$$

Since $(p+\epsilon)s(1,x) + (1-p-\epsilon)s(0,x)$ is minimized at $x = p+\epsilon$ by Definition 2(a), the last term in square brackets is negative. Hence

$$\lim_{\epsilon \to 0} \frac{1}{\epsilon}\left(\varphi(p+\epsilon) - \varphi(p)\right) \geqslant \psi(p) ,$$

and similarly one shows

$$\lim_{\epsilon \to 0} \frac{1}{\epsilon}\left(\varphi(p) - \varphi(p-\epsilon)\right) \leqslant \psi(p) .$$

Since $\psi$ is continuous by Definition 2(b), this shows that $\varphi$ is differentiable, and hence $\psi = \varphi'$. This proves Eq. (2). Continuity of $\varphi$ up to the boundary of $[0,1]$ follows from continuity of $s$ and Lemma 1.

To prove the converse, first note that if $\varphi$ is bounded and convex on $(0,1)$, it can be extended to a continuous function on $[0,1]$, as shown in Lemma 1. Because of strict convexity of $\varphi$ we have, for $p \in [0,1]$ and $0 < x < 1$,

$$ps(1,x) + (1-p)s(0,x) = -\varphi(x) - \varphi'(x)(p-x) \geqslant -\varphi(p) , \tag{7}$$

with equality if and only if $x = p$.

It remains to show that the same is true for $x \in \{0,1\}$. Consider first the case $x = 0$. We have to show that $ps(1,0) + (1-p)s(0,0) > -\varphi(p)$ for $p > 0$. By continuity of $s$, Eq. (2) and Lemma 1, we have $s(1,0) = -\varphi(0) - \lim_{p\to 0} \varphi'(p)$, while $s(0,0) = -\varphi(0)$. If $\lim_{p\to 0} \varphi'(p) = -\infty$, the result is immediate. If $\varphi'(0) := \lim_{p \to 1} \varphi'(p)$ is finite, we have $-\varphi(0) - p\varphi'(0) > -\varphi(p)$ again by strict convexity of $\varphi$.

Likewise, one shows that $ps(1,1) + (1-p)s(0,1) > -\varphi(p)$ for $p < 1$. This finishes the proof that $s$ is a proper scoring rule. □

*Proof of Proposition 3.* For fixed $\mathbf{x} \in \mathsf{C}$, the function $\mathbf{y} \mapsto d_\Phi(\mathbf{y}, \mathbf{x})$ is strictly convex, and hence achieves a unique minimum at a point $\boldsymbol{\pi}_\mathbf{x}$ in the convex, closed set $\mathsf{Z}$.

Let $\mathbf{y} \in \mathsf{Z}$. For $0 \leqslant \epsilon \leqslant 1$, $(1-\epsilon)\boldsymbol{\pi}_\mathbf{x} + \epsilon\mathbf{y} \in \mathsf{Z}$, and hence $d_\Phi((1-\epsilon)\boldsymbol{\pi}_\mathbf{x} + \epsilon\mathbf{y}, \mathbf{x}) - d_\Phi(\boldsymbol{\pi}_\mathbf{x}, \mathbf{x}) \geqslant 0$ by the definition of $\boldsymbol{\pi}_\mathbf{x}$. Since $d_\Phi$ is differentiable in the first argument, we can divide by $\epsilon$ and let $\epsilon \to 0$ to obtain

$$0 \leqslant \lim_{\epsilon \to 0} \frac{1}{\epsilon} \left( d_\Phi((1-\epsilon)\boldsymbol{\pi}_\mathbf{x} + \epsilon\mathbf{y}, \mathbf{x}) - d_\Phi(\boldsymbol{\pi}_\mathbf{x}, \mathbf{x}) \right) = (\nabla\Phi(\boldsymbol{\pi}_\mathbf{x}) - \nabla\Phi(\mathbf{x})) \cdot (\mathbf{y} - \boldsymbol{\pi}_\mathbf{x})\,.$$

The fact that

$$d_\Phi(\mathbf{y}, \mathbf{x}) - d_\Phi(\boldsymbol{\pi}_\mathbf{x}, \mathbf{x}) - d_\Phi(\mathbf{y}, \boldsymbol{\pi}_\mathbf{x}) = (\nabla\Phi(\boldsymbol{\pi}_\mathbf{x}) - \nabla\Phi(\mathbf{x})) \cdot (\mathbf{y} - \boldsymbol{\pi}_\mathbf{x})$$

proves the claim. □

# 6 Proof of Theorem 1

The main idea of the proof is more apparent when $s$ is bounded. So we consider this case on its own before allowing $s$ to reach $+\infty$.

**Bounded Case.**

Suppose $s$ is bounded. In this case, the derivative of the corresponding $\varphi$ from Eq. (2) in Proposition 2 is continuous and bounded all the way up to the boundary of $[0, 1]$.

Let $\mathbf{f} \in [0, 1]^n$ be a forecast and, for $\omega \in \Omega$, let $\boldsymbol{\nu}_\omega \in \mathsf{V}$ be the vector with components $C_{E_i}(\omega)$. Let $\Phi(\mathbf{x}) = \sum_{i=1}^{n} \varphi(x_i)$. Then

$$\begin{aligned}
P_s(\omega, \mathbf{f}) &= \sum_{i=1}^{n} s(C_{E_i}(\omega), f_i) \quad \text{[Definition 3]} \\
&= \sum_{i=1}^{n} -\varphi(f_i) - \varphi'(f_i)(C_{E_i}(\omega) - f_i) \quad \text{[Proposition 2]} \\
&= d_\Phi(\boldsymbol{\nu}_\omega, \mathbf{f}) - \sum_{i=1}^{n} \varphi(C_{E_i}(\omega)) \quad \text{[Definition 6]} \\
&= d_\Phi(\boldsymbol{\nu}_\omega, \mathbf{f}) + \sum_{i=1}^{n} s(C_{E_i}(\omega), C_{E_i}(\omega)) \quad \text{[Equation 3].}
\end{aligned} \tag{8}$$

Now assume that $\mathbf{f}$ is incoherent which, by Proposition 1, means that $\mathbf{f} \notin \mathit{conv}(\mathsf{V})$. According to Eq. (4) of Proposition 3, there exists a $\mathbf{g} \in \mathit{conv}(\mathsf{V})$, namely the projection of $f$ onto *conv*$(V)$, such that $d_\Phi(y, g) \leqslant d_\Phi(y, f) - d_\Phi(g, f)$ for all $y \in$ *conv*$(V)$ and hence, in particular, for $y \in V$. Since $d_\Phi(g, f) > 0$ this proves part $(b)$ of Theorem 1.

To prove part $(a)$ first note that weak dominance of $f$ by $g$ means that $d_\Phi(\nu_\omega, g) \leqslant d_\Phi(\nu_\omega, f)$ for all $\nu_\omega \in V$, by Eq. (8). In this case, $d_\Phi(y, g) \leqslant d_\Phi(y, f)$ for all $y \in$ *conv*$(V)$, since $d_\Phi(y, g) - d_\Phi(y, f)$ depends linearly on $y$. If $f$ is coherent, $f \in$ *conv*$(V)$ by Proposition 1, and hence $d_\Phi(f, g) \leqslant d_\Phi(f, f) = 0$. This implies that $g = f$.

**Unbounded Case.**

Next, consider the case when $s$ is unbounded. In this case, the derivative of the corresponding $\varphi$ from Proposition 2 diverges either at 0 or 1, or at both values, and hence we can not directly apply Proposition 3. Eq. (8) is still valid, with both sides of the equation possibly being $+\infty$. However, if $f$ lies either in the interior of $[0, 1]^n$, or on a point on the boundary where the derivative of $\Phi(x) = \sum_i \varphi(x_i)$ does not diverge, an examination of the proof of Proposition 3 shows that the result still applies, as we show now.

If $\nabla\Phi(f)$ is finite, the minimum of $\Phi(y) - \nabla\Phi(f) \cdot y$ over $y \in$ *conv*$(V)$ is uniquely attained at some $g \in$ *conv*$(V)$. Moreover, $\nabla\Phi(g)$ is necessarily finite. Repeating the argument in the proof of Proposition 3 shows that $d_\Phi(y, g) \leqslant d_\Phi(y, f) - d_\Phi(g, f)$ for any $y \in$ *conv*$(V)$, which is the desired inequality needed in the proof of Theorem 1$(b)$. We are thus left with the case in which $f$ lies on an $(n-1)$ dimensional face of $[0, 1]^n$ where the normal derivative diverges. Consider first the case $n = 1$. Then either $V = \{0, 1\}$, in which case $f$ is coherent, or $V = \{0\}$ or $\{1\}$, in which case it is clear that the unique coherent vector $g \in V$ strongly dominates $f$.

We now proceed by induction on the dimension $n$ of the forecast $f$. In the $(n-1)$ dimensional hypercube, either $f$ lies inside or on a point of the boundary where the normal derivative of $\Phi$ is finite, in which case we have just argued that there exists a $\tilde{g}$ that is coherent and satisfies $P_s(\omega, \tilde{g}) < P_s(\omega, f)$ for all $\omega$ such that $\nu_\omega$ lies in the $(n-1)$ dimensional face. In the other case, the induction hypothesis implies that we can find such a $\tilde{g}$. Note that for all the other $\omega$, $P_s(\omega, \tilde{g}) = P_s(\omega, f) = \infty$. Now simply pick an $0 < \epsilon < 1$ and choose $g_\epsilon = (1-\epsilon)\tilde{g} + \epsilon l^{-1} \sum_{i=1}^{l} \nu_i$, where the $\nu_i$ denote all the $l$ elements of $V$ outside the $(n-1)$-dimensional hypercube. Then $P_s(\omega, g_\epsilon) < \infty$ for all $\omega$ and also, using Lemma 1, $\lim_{\epsilon\to 0} P_s(\omega, g_\epsilon) = P_s(\omega, \tilde{g})$. Hence we can choose $\epsilon$ small enough to conclude that $P_s(\omega, g_\epsilon) < P_s(\omega, f)$ for all $\omega \in \Omega$. This finishes the proof of part $(b)$ in the general case of unbounded $s$.

To prove part $(a)$ in the general case, we note that if $f = \sum_i a_i \nu_i$ for $\nu_i \in V$ and $a_i > 0$, then necessarily $d_\Phi(\nu_i, f) < \infty$. That is, any coherent $f$ is a convex combination of $\nu_i \in V$ such that $d_\Phi(\nu_i, f) < \infty$. This follows from the fact that a component of $f$

can be 0 only if this component is 0 for all the $\nu_i$'s. The same is true for the value 1. But the $d_\Phi(\nu, f)$ can be infinite only if some component of $f$ is 0 and the corresponding one for $\nu$ is 1, or vice versa.

Since $d_\Phi(\nu_i, f) < \infty$ for the $\nu_i$ in question, also $d_\Phi(\nu_i, g) < \infty$ by Eq. (8) and the assumption that $f$ is weakly dominated by $g$. Moreover, $d_\Phi(\nu_i, g) - d_\Phi(\nu_i, f) \leqslant 0$. But $\sum_i a_i(d_\Phi(\nu_i, g) - d_\Phi(\nu_i, f)) = d_\Phi(f, g) \geqslant 0$, hence $f = g$. □

# 7 Generalizations

## 7.1 Penalty functions

Theorem 1 holds for a larger class of penalty functions. In fact, one can use different proper scoring rules for every event, and replace (1) by

$$P_s(\omega, f) = \sum_{i \leqslant n} s_i(C_{E_i}(\omega), f_i)\,,$$

where the $s_i$ are possibly distinct proper scoring rules. In this way, forecasts for some events can be penalized differently than others. The relevant Bregman divergence in this case is given by $\Phi(x) = \sum_i \varphi_i(x_i)$, where $\varphi_i$ is determined by $s_i$ via (2). Proof of this generalization closely follows the argument given above, so it is omitted. Additionally, by considering more general convex functions $\Phi$ our argument generalizes to certain non-additive penalties.

## 7.2 Generalized scoring rules

### 7.2.1 Non-uniqueness

If one relaxes the condition of *unique* minimization in Definition 2(a), a weaker form of Theorem 1 still holds. Namely, for any incoherent forecast $f$ there exists a coherent forecast $g$ that weakly dominates $f$. Strong dominance will not hold in general, as the example of $s(i, x) \equiv 0$ shows.

Proposition 2 also holds in this generalized case, but the function $\varphi$ need not be *strictly* convex. Likewise, Proposition 3 can be generalized to merely convex (not necessarily strictly convex) $\Phi$ but in this case the projection $\pi_x$ need not be unique. Eq. (4) remains valid.

### 7.2.2 Discontinuity

A generalization that is more interesting mathematically is to discontinuous scoring rules. Proposition 2 can be generalized to scoring rules that satisfy neither the continuity condition in Definition 2 nor unique minimization. (This is also shown in Gneiting and Raftery, 2007).

**Proposition 4.** Let $s : \{0,1\} \times [0,1] \to [0,\infty]$ satisfy

$$ps(1,x) + (1-p)s(0,x) \geqslant ps(1,p) + (1-p)s(0,p) \quad \forall x, p \in [0,1]\,. \tag{9}$$

Then the function $\varphi : [0,1] \mapsto \mathfrak{R}$ defined by $\varphi(x) = -xs(1,x) - (1-x)s(0,x)$ is bounded and convex. Moreover, there exists a monotone non-decreasing function $\psi : [0,1] \mapsto \mathfrak{R} \cup \{\pm\infty\}$, with the property that

$$\psi(x) \geqslant \lim_{\epsilon \to 0} \frac{1}{\epsilon} \left(\varphi(x) - \varphi(x-\epsilon)\right) \quad \forall x \in (0,1]\,, \tag{10}$$

$$\psi(x) \leqslant \lim_{\epsilon \to 0} \frac{1}{\epsilon} \left(\varphi(x+\epsilon) - \varphi(x)\right) \quad \forall x \in [0,1)\,, \tag{11}$$

such that

$$s(i,x) = -\varphi(x) - \psi(x)(i-x) \quad \forall x \in (0,1)\,. \tag{12}$$

Function $\varphi$ is strictly convex if and only if the inequality (9) is strict for $x \neq p$.

Conversely, if $s$ is of the form (12), with $\varphi$ bounded and convex and $\psi$ satisfying (10)–(11), then $s$ satisfies (9).

It is a fact (Hardy et al., 1934) that every convex function $\varphi$ on $[0,1]$ is continuous on $(0,1)$ and has a right and left derivative, $\psi_R$ and $\psi_L$ (defined by the right sides of (11) and (10), respectively) at every point (except the endpoints, where it has only a right or left derivative, respectively). Both $\psi_R$ and $\psi_L$ are non-decreasing functions, and $\psi_L(x) \leqslant \psi_R(x)$ for all $x \in (0,1)$. Except for countably many points, $\psi_L(x) = \psi_R(x)$, i.e., $\varphi$ is differentiable. Eqs. (10)–(11) say that $\psi_L(x) \leqslant \psi(x) \leqslant \psi_R(x)$. The concept of subgradient, well known in convex analysis (Rockafellar, 1970), plays the role of derivative for non-differentiable convex functions.

Note that although $s(0,x)$ and $s(1,x)$ may be discontinuous, the combination $\varphi(x) = -xs(1,x) - (1-x)s(0,x)$ is continuous. Hence, if $s(0,x)$ jumps up at a point $x$, $s(1,x)$ has to jump down by an amount proportional to $(1-x)/x$.

The proof of Proposition 4 is virtually the same as the proof of Proposition 2, so we omit it.

### 7.3 Open question

Whether Theorem 1 holds for this generalized notion of a discontinuous scoring rule remains open. The proof of Theorem 1 given here does not extend to the discontinuous case, since for inequality (4) to hold, differentiability of $\Phi$ is necessary, in general.

## References

A. Banerjee, X. Guo, and H. Wang. On the optimality of conditional expectation as a Bregman predictor. *IEEE Transactions on Information Theory*, 51(7):26642669, 2005.

J. M. Bernardo and A. F. M. Smith. *Bayesian Theory.* John Wiley & Sons, West Sussex, England, 1994.

L. M. Bregman. The relaxation method of finding a common point of convex sets andits application to the solution of problems in convex programming. *U. S. S. R. Computational Mathematics and Mathematical Physics*, 78(384):200–217, 1967.

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

Y. Censor and S. A. Zenios. *Parallel Optimization: Theory, Algorithms, and Applications.* Oxford University Press, 1997.

B. de Finetti. *Theory of Probability*, volume 1. John Wiley and Sons, New York, NY, 1974.

T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, March 2007.

I. J. Good. Rational decisions. *Journal of the Royal Statistical Society*, 14:107–114, 1952.

G. H. Hardy, J. E. Littlewood, and G. Pólya. *Inequalities.* Cambridge University Press, 1934.

J. M. Joyce. A nonpragmatic vindication of probabilism. *Philosophy of Science*, 65: 575603, 1998.

D. V. Lindley. Scoring rules and the inevitability of probability. *International Statistical Review*, 50:1–26, 1982.

R. T. Rockafellar. *Convex Analysis.* Princeton University Press, 1970.

L. J. Savage. Elicitation of personal probabilities and expectations. *Journal of the Americal Statistical Association*, 66(336):783–801, 1971.

R. Selten. Axiomatic characterization of the quadratic scoring rule. *Experimental Economics*, 1:43–62, 1998.

B. Skyrms. *Choice & Chance: An Introduction to Inductive Logic.* Wadsworth, Belmont CA, 2000.

Joel Predd
Rand Corporation
4570 Fifth Avenue, Suite 600
Pittsburgh, PA 15213
jpredd@rand.org

Robert Seiringer
Dept. of Physics
Princeton University
Princeton NJ 08540
rseiring@princeton.edu

Elliott Lieb
Depts. of Mathematics and Physics
Princeton University
Princeton NJ 08540
lieb@princeton.edu

Daniel Osherson
Dept. of Psychology
Princeton University
Princeton NJ 08540
osherson@princeton.edu

Vincent Poor
Dept. of Electrical Engineering
Princeton University
Princeton NJ 08540
poor@princeton.edu

Sanjeev Kulkarni
Dept. of Electrical Engineering
Princeton University
Princeton NJ 08540
kulkarni@princeton.edu