Evaluating Imprecise Forecasts

Jason Konek JASON.KONEK@BRISTOL.AC.UK
Department of Philosophy, University of Bristol, UK

Abstract

This paper will introduce a new class of IP scoring rules for sets of almost desirable gambles. A set of almost desirable gambles $\mathcal{D}$ is evaluable for what might be called generalised type 1 and type 2 error. Generalised type 1 error is roughly a matter of the extent to which $\mathcal{D}$ encodes false judgments of desirability. Generalised type 2 error is roughly a matter of the extent to which $\mathcal{D}$ fails to encode true judgments of desirability. IP scoring rules are penalty functions that average these two types of error. To demonstrate the viability of IP scoring rules, we must show that for any coherent $\mathcal{D}$ you might choose, we can construct an IP scoring rule that renders it admissible. Moreover, every other admissible model relative to that scoring rule is also coherent. This paper makes progress toward that goal. We will also compare the class of scoring rules developed here with the results by Seidenfeld, Schervish, and Kadane from 2012,which establish that there is no strictly proper, continuous real-valued scoring rule for lower and upper probability forecasts.

Keywords: scoring rules, accuracy, forecasting, lower previsions, closed convex sets of probabilities, sets of almost desirable gambles

1. Introduction

Policy-makers and stakeholders often rely on expert forecasts to inform their decision-making. For example, during the COVID-19 pandemic, the US Center for Disease Control funded the Delphi group to run the COVID-19 Forecast Hub and invited modelling teams from around the world to submit forecasts of case numbers, hospitalizations, and death counts. Given the uncertainty around these variables, modellers were not asked to submit a single, real-valued forecast for each one. Rather, they were asked to issue predictive intervals which were treated by assessors at the Delphi group as lower and upper previsions—*imprecise forecasts*.

At the beginning of the pandemic, the Delphi group constructed a single ensemble forecast by averaging the lower and upper previsions of all eligible models in COVID-19 Forecast Hub. By November 2021, they transitioned to building an ensemble from the best performing models over the previous 12 week period. To do this, they needed a method for evaluating the accuracy of past forecasts. They chose to score the lower/upper prevision pair $\langle l, u\rangle$ for variable X by the Interval Score (at a specified α level).

$$IS_\alpha(l,u,x) = (u-l)+\frac{2}{\alpha}(l-x)\mathbb{1}(l>x)+\frac{2}{\alpha}(x-u)\mathbb{1}(x>u)$$

The interval score penalises $\langle l, u\rangle$ for imprecision (measured by the $(u-l)$ term), but also penalises $\langle l, u\rangle$ if the true value x of X falls outside the interval $[l, u]$ (the further outside, the worse). Different lower/upper prevision pairs were scored at different α levels (in a way that incentivised modellers to report central predictive intervals as lower/upper previsions) and these scores were then averaged. The resulting average is the Weighted Interval Score (WIS).

To build an ensemble forecast, the Delphi group selected the 10 models with the best performance according to the WIS over the prior 12 week period. Component models were assigned weights based on their relative WIS during those 12 weeks. Models with a stronger record of accuracy received higher weight.

IP scoring rules like the WIS can serve a variety of theoretical and practical functions. They are useful for incentivising experts to report IP forecasts and deciding how to aggregate those forecasts, *e.g.*, to produce a single ensemble forecast for use in policy and decision-making. They are also potentially useful for incentivising traders to produce IP forecasts in prediction markets, for evaluating and improving IP expert systems in medicine, and for training neural net classifiers to produce imprecise classification probabilities.

The aim of this paper is to provide a general method for constructing IP scoring rules for a range of equivalent imprecise probability models: coherent lower previsions, nonempty closed convex sets of probability measures, and coherent sets of almost-desirable gambles.

2. Framework

We will focus our attention, in the first instance, on developing IP scoring rules for epigraphical sets of almost desirable gambles. We can then score coherent lower previsions and nonempty closed convex sets of probability measures by scoring the equivalent coherent set of almost desirable gambles.

Let $\Omega = \{\omega_1, \ldots, \omega_n\}$ be a finite possibility space.

A gamble $g : \Omega \to \mathbb{R}$ is an uncertain reward which pays out in linear utility. We will treat them as elements $g = \langle g_1, \dots, g_n \rangle$ of $\mathbb{R}^n$.

A set $\mathcal{D} \subseteq \mathbb{R}^n$ is a **coherent** set of almost desirable gambles iff it satisfies:

AD1. If $g < 0$ then $g \notin \mathcal{D}$ (where $g < 0 \Leftrightarrow g_i < 0$ for all $i \leqslant n$)

AD2. If $g \geqslant 0$ then $g \in \mathcal{D}$ (where $g \geqslant 0 \Leftrightarrow g_i \geqslant 0$ for all $i \leqslant n$)

AD3. If $g \in \mathcal{D}$ and $\lambda > 0$ then $\lambda g \in \mathcal{D}$

AD4. If $f, g \in \mathcal{D}$ then $f + g \in \mathcal{D}$

AD5. If $g + \epsilon \in \mathcal{D}$ for all $\epsilon > 0$ then $g \in \mathcal{D}$

The **epigraph** of a function $b : \mathbb{R}^{n-1} \to [-\infty, \infty]$ is

$$\mathcal{D}_b = \{\langle g_1, \dots, g_n \rangle \,|\, g_n \geqslant b(g_1, \dots, g_{n-1})\} \subseteq \mathbb{R}^n$$

Every coherent set of almost-desirable gambles $\mathcal{D}$ is epigraphical.[1] But many epigraphical sets of almost desirable gambles are not coherent. Indeed,

Proposition 1 *For any $b : \mathbb{R}^{n-1} \to [-\infty, \infty]$, $\mathcal{D}_b \subseteq \mathbb{R}^n$ is a coherent set of almost desirable gambles if and only if*

E1. If $g \geqslant 0$ then $b(g) \leqslant 0$

E2. If $g \leqslant 0$ then $b(g) \geqslant 0$

E3. If $\lambda > 0$ then $b(\lambda g) = \lambda b(g)$

E4. $b(f + g) \leqslant b(f) + b(g)$

We will score an epigraphical set of almost desirable gambles $\mathcal{D}$ at a state of the world ω_i by comparing it with the "ideal set of almost desirable gambles" at ω_i. The **ideal set of almost desirable gambles** if ω_i is the true state of the world is given by

$$\mathcal{D}_i = \{g | g_i \geqslant 0\} \subseteq \mathbb{R}^n$$

$\mathcal{D}_i$ contains all and only the gambles that are *in fact* almost desirable at ω_i.

For any epigraphical set $\mathcal{D} \subseteq \mathbb{R}^n$ of almost desirable gambles—coherent or not—let $\mathcal{E}_i^1 = \mathcal{D} \setminus \mathcal{D}_i$ and $\mathcal{E}_i^2 = \mathcal{D}_i \setminus \mathcal{D}$. $\mathcal{E}_i^1$ is the set of *type 1 errors* that $\mathcal{D}$ commits at ω_i. If $\mathcal{D}$ says that g is almost desirable but it is not—*i.e.*, if $g \in \mathcal{D}$ but $g_i < 0$ so that $g \notin \mathcal{D}_i$—then $\mathcal{D}$ commits a type 1 error. It says something false about g (that it is almost desirable). $\mathcal{E}_i^1$ collects up all of the gambles that $\mathcal{D}$ falsely characterises as almost desirable. Likewise $\mathcal{E}_i^2$ is the set of *type 2 errors* that $\mathcal{D}$ commits at ω_i. If g is in fact almost desirable but $\mathcal{D}$ fails to say so—*i.e.*, $g_i \geqslant 0$ so that $g \in \mathcal{D}_i$ but $g \notin \mathcal{D}$—then $\mathcal{D}$ commits a type 2 error. It fails to say something true about g (that it is almost desirable). $\mathcal{E}_i^2$ collects up all of the gambles that $\mathcal{D}$ fails to (truly) characterise as almost desirable.

Let $\mathcal{E}_i$ be the total set of gambles that $\mathcal{D}$ mischaracterizes at ω_i: $\mathcal{E}_i = \mathcal{E}_i^1 \bigcup \mathcal{E}_i^2$. Call this $\mathcal{D}$'s *error set*. The central tenet of our approach to IP scoring rules is this:

Inaccuracy is a measure of error.

Formally, we capture this by saying that the inaccuracy of $\mathcal{D}$ at ω_i, $\mathcal{I}(\mathcal{D}, \omega_i)$, is the measure of $\mathcal{E}_i$ according to an appropriate measure ν_i:

$$\mathcal{I}(\mathcal{D}, \omega_i) = \mathcal{I}_i(\mathcal{D}) = \nu_i(\mathcal{E}_i)$$

Intuitively, $\nu_i(\mathcal{E}_i)$ captures something like the size of the error set $\mathcal{E}_i$. We assume that ν_i is a measure on the Borel σ-algebra $\mathfrak{B}(\mathbb{R}^n)$ that is (i) finite, *i.e.*, $\nu_i(\mathbb{R}^n) < \infty$, and (ii) absolutely continuous with respect to the product Lebesgue measure μ. In that case, the Radon–Nikodym theorem guarantees that there is some measurable function $\phi_i : \mathbb{R}^n \to \mathbb{R}$ such that

$$\mathcal{I}_i(\mathcal{D}) = \int_{\mathcal{E}_i} |\phi_i| \,\mathrm{d}\mu$$

We will also assume that ϕ_i can be chosen to be continuously differentiable [10].

Let $\mathcal{F}_i(D) = \int_{\mathcal{E}_i^1} |\phi_i| \,\mathrm{d}\mu$ and $\mathcal{S}_i(D) = \int_{\mathcal{E}_i^2} |\phi_i| \,\mathrm{d}\mu$ so that

$$\mathcal{I}_i(\mathcal{D}) = \mathcal{F}_i(D) + \mathcal{S}_i(D)$$

We can think of $|\phi_i(g)|$ as a penalty that you would receive for mischaracterizing g. If $g_i < 0$, then $|\phi_i(g)|$ is the penalty you would receive for falsely characterizing g as almost desirable. If $g_i \geqslant 0$, then $|\phi_i(g)|$ is the penalty you would receive for failing to truly characterize g as almost desirable. For $\mathcal{I}$ to count as a measure of inaccuracy, we will insist that it satisfy at least the following two axioms:

P_1. $\phi_i(g_1, \dots, g_n)$ is (at least weakly) increasing in g_i

P_2. $\phi_i(g_1, \dots, g_{i-1}, 0, g_{i+1}, \dots, g_n) = 0$

P1 and P2 jointly capture the idea that all else equal, accepting a bigger loss is a bigger type 1 error, and leaving more utility on the table is a bigger type 2 error.

The central task in developing the foundations of IP scoring rules is to jointly axiomatize reasonable ν_i and ϕ_i. Rather than plumping for some plausible sounding axioms,

[1] Let $b(g_1, \dots, g_{n-1}) = g_n$ if $g_n = \inf\{z | \underline{P}(g_1, \dots, g_{n-1}, z) \geqslant 0\}$ and $b(g_1, \dots, g_{n-1}) = -\infty$ if $\inf\{z | \underline{P}(g_1, \dots, g_{n-1}, z) \geqslant 0\}$ does not exist, where $\underline{P}$ is the coherent lower prevision that is equivalent to $\mathcal{D}$.

we will attempt to tease axioms on ν_i and ϕ_i out from our coherence axioms AD1-AD5. We would like incoherent sets of almost desirable gambles to be inadmissible, *i.e.*, dominated relative to $\mathcal{I}$. We would also like our axioms on ν_i and ϕ_i to be flexible enough so that for any coherent set of almost desirable gambles, there is some $\mathcal{I}$ relative to which it is admissible. We will attempt to pin down what ν_i and ϕ_i must look like in order to make this so. The investigation in this paper will make progress toward that end, but will not get us all of the way there.

3. Strictly Proper Scoring Rules

It will prove instructive to briefly consider the relationship between the IP scoring rules outlined in section 1 and *strictly proper scoring rules*.

Let $\mathcal{F}$ be the power set of Ω and $c : \mathcal{F} \to \mathbb{R}$ be a (not necessarily probabilistic) assignment of precise forecasts to the events in $\mathcal{F}$. Let C be the space of all such assignments. A scoring rule $\mathcal{I} : C \times \Omega \to \mathbb{R}_{\geqslant 0}$ is **strictly proper** iff

$$\sum_{\omega \in \Omega} p(\omega)\mathcal{I}(p, \omega) < \sum_{\omega \in \Omega} p(\omega)\mathcal{I}(c, \omega)$$

for any probability function $p \in C$ and any $c \neq p$.[2] Strictly proper scoring rules have been studied extensively, *e.g.*, their foundations [2, 4, 9], in elicitation [12], their connection to guidance value in binary decision problems [13], and their connection to coherence [11, 7]. Popular strictly proper scoring rules include:

- **Brier Score**: $\mathcal{I}(c, \omega) = \sum_{X \in \mathcal{F}} (\mathbb{1}_X(\omega) - c(X))^2$
- **Log Score**: $\mathcal{I}(c, \omega) = \sum_{X \in \mathcal{F}} -log(|1 - \mathbb{1}_X(\omega) - c(X)|)$
- **Spherical Score**:

$$\mathcal{I}(c, \omega) = \sum_{X \in \mathcal{F}} \left(1 - \frac{|1 - \mathbb{1}_X(\omega) - c(X)|}{\sqrt{c(X)^2 + (1 - c(X))^2}}\right)$$

Strictly proper scoring rules are a special case of the scoring rules in section 1 under special assumptions. To see this, first choose a probability mass function $p : \Omega \to \mathbb{R}$. Let

$$\mathcal{D}_p = \{g | p.g \geqslant 0\}$$

$\mathcal{D}_p$ is the (coherent) set of almost desirable gambles associated with p. Our question is: under what conditions do we have

$$\sum_{\omega \in \Omega} p(\omega)\mathcal{I}(\mathcal{D}_p, \omega) < \sum_{\omega \in \Omega} p(\omega)\mathcal{I}(\mathcal{D}, \omega)$$

for any $\mathcal{D} \neq \mathcal{D}_p$?

Suppose that our IP scoring rule

$$\mathcal{I}(\mathcal{D}, \omega_i) = \mathcal{I}_i(\mathcal{D}) = \nu_i(\mathcal{E}_i) = \int_{\mathcal{E}_i} |\phi_i| \, d\mu$$

is given by measures $\nu_1, \ldots, \nu_n$ that satisfy some (unnecessarily) restrictive assumptions.

SP1. $\phi_i(\lambda g) = \lambda \phi_i(g)$ for any $\lambda > 0$ and g in measurable X

SP2. $\nu_i(\mathcal{E}_i) = \nu_i(\mathcal{E}_i^*)$ for any $\mathcal{E}_i^*$ s.t.

$$\mathcal{E}_i^* = \{\langle x_1, \ldots, x_{i-i}, g_i, x_{i+1}, \ldots, x_n \rangle \mid g \in \mathcal{E}_i, x_1, \ldots, x_n \in \mathbb{R}\}$$

SP3. $\nu_i(\mathcal{E}_i) = \nu_j(\mathcal{E}_j^\dagger)$ where $\mathcal{E}_j^\dagger$ is the result of permuting the i^{th} and j^{th} component of any $g \in \mathcal{E}_i$, *i.e.*,

$$\mathcal{E}_j^\dagger = \left\{g^\dagger \middle| g \in \mathcal{E}_i, g_i^\dagger = g_j, g_j^\dagger = g_i, g_k^\dagger = g_k \text{ for all } k \neq i, j\right\}$$

SP1 ensures that the penalty you receive for mischaracterizing a gamble g scales linearly with its (linear utility) payout. SP2 ensures that $\mathcal{I}_i(\mathcal{D})$ depends only on the values that the gambles $g \in \mathcal{E}_i$ take at ω_i. Whether those gambles *would* have yielded huge gains (or losses) in other states does not affect your *actual* degree of type 1 and type 2 error. Given SP2, SP3 ensures that $\mathcal{I}_i(\mathcal{D}) = \mathcal{I}_j(\mathcal{D})$ whenever $\mathcal{E}_i$ and $\mathcal{E}_j$ make equivalent mistakes at ω_i and ω_j, respectively.

SP1 states that ϕ_i is positive homogenous of degree 1 (for all $i \leqslant n$). Euler's homogeneous function theorem then guarantees that

$$\phi_i(g_1, \ldots, g_n) = \sum_{j \leqslant n} g_j \frac{\partial \phi_i}{\partial g_j}(g)$$

SP2 further implies that

$$\phi_i(g_1, \ldots, g_n) = g_i \frac{\partial \phi_i}{\partial g_i}(g)$$

But this is true iff $\phi_i(g) = c_i g_i$ for some $c_i > 0$. SP3 further implies that $c_i = c_j$ for all $i, j \leqslant n$. In that case, it is straightforward to show that $\mathcal{I}$ is strictly proper.

Proposition 2 *If there is some $c > 0$ such that for all $i \leqslant n$*

$$\mathcal{I}_i(\mathcal{D}) = \int_{\mathcal{E}_i} |c g_i| \, d\mu$$

Then for any probability mass function $p : \Omega \to \mathbb{R}$ and any $\mathcal{D} \neq \mathcal{D}_p$

$$\sum_{i \leqslant n} p_i \mathcal{I}_i(\mathcal{D}_p) < \sum_{i \leqslant n} p_i \mathcal{I}_i(\mathcal{D})$$

unless both $\mathcal{D} \setminus \mathcal{D}_p$ and $\mathcal{D}_p \setminus \mathcal{D}$ are sets of measure zero.

[2] Some authors reserve the term "scoring rule" for the summands of additive loss functions.

Proof We must show that

$$\begin{aligned}
&\sum_{i\leqslant n} p_i \left[\mathcal{I}_i(\mathcal{D}) - \mathcal{I}_i(\mathcal{D}_p)\right] \\
&\quad = \sum_{i\leqslant n} p_i \left[\mathcal{F}_i(\mathcal{D}) + \mathcal{S}_i(\mathcal{D}) - \mathcal{F}_i(\mathcal{D}_p) - \mathcal{S}_i(\mathcal{D}_p)\right] \\
&\quad = \sum_{i\leqslant n} p_i \left[\mathcal{F}_i(\mathcal{D}) - \mathcal{F}_i(\mathcal{D}_p)\right] \\
&\qquad + \sum_{i\leqslant n} p_i \left[\mathcal{S}_i(\mathcal{D}) - \mathcal{S}_i(\mathcal{D}_p)\right] \\
&\quad > 0
\end{aligned}$$

Firstly, note that

$$\begin{aligned}
&\mathcal{F}_i(\mathcal{D}) - \mathcal{F}_i(\mathcal{D}_p) \\
&\quad = \int_{\mathcal{D}\setminus\mathcal{D}_i} -cg_i \,\mathrm{d}\mu - \int_{\mathcal{D}_p\setminus\mathcal{D}_i} -cg_i \,\mathrm{d}\mu \\
&\quad = \int_{\mathcal{D}\setminus(\mathcal{D}_p\cup\mathcal{D}_i)} -cg_i \,\mathrm{d}\mu - \int_{\mathcal{D}_p\setminus(\mathcal{D}\cup\mathcal{D}_i)} -cg_i \,\mathrm{d}\mu
\end{aligned}$$

Likewise,

$$\begin{aligned}
&\mathcal{S}_i(\mathcal{D}) - \mathcal{S}_i(\mathcal{D}_p) \\
&\quad = \int_{\mathcal{D}_i\setminus\mathcal{D}} cg_i \,\mathrm{d}\mu - \int_{\mathcal{D}_i\setminus\mathcal{D}_p} cg_i \,\mathrm{d}\mu \\
&\quad = \int_{\mathcal{D}_p\setminus(\mathcal{D}\cup\mathcal{D}_i^C)} cg_i \,\mathrm{d}\mu - \int_{\mathcal{D}\setminus(\mathcal{D}_p\cup\mathcal{D}_i^C)} cg_i \,\mathrm{d}\mu
\end{aligned}$$

In that case

$$\begin{aligned}
&\sum_{i\leqslant n} p_i \left[\mathcal{F}_i(\mathcal{D}) - \mathcal{F}_i(\mathcal{D}_p)\right] + \sum_{i\leqslant n} p_i \left[\mathcal{S}_i(\mathcal{D}) - \mathcal{S}_i(\mathcal{D}_p)\right] \\
&= \sum_{i\leqslant n} p_i \left[\int_{\mathcal{D}\setminus(\mathcal{D}_p\cup\mathcal{D}_i)} -cg_i \,\mathrm{d}\mu - \int_{\mathcal{D}_p\setminus(\mathcal{D}\cup\mathcal{D}_i)} -cg_i \,\mathrm{d}\mu\right] \\
&\quad + \sum_{i\leqslant n} p_i \left[\int_{\mathcal{D}_p\setminus(\mathcal{D}\cup\mathcal{D}_i^C)} cg_i \,\mathrm{d}\mu - \int_{\mathcal{D}\setminus(\mathcal{D}_p\cup\mathcal{D}_i^C)} cg_i \,\mathrm{d}\mu\right] \\
&= \sum_{i\leqslant n} p_i \left[\int_{\mathcal{D}_p\setminus(\mathcal{D}\cup\mathcal{D}_i^C)} cg_i \,\mathrm{d}\mu + \int_{\mathcal{D}_p\setminus(\mathcal{D}\cup\mathcal{D}_i)} cg_i \,\mathrm{d}\mu\right] \\
&\quad - \sum_{i\leqslant n} p_i \left[\int_{\mathcal{D}\setminus(\mathcal{D}_p\cup\mathcal{D}_i)} cg_i \,\mathrm{d}\mu + \int_{\mathcal{D}\setminus(\mathcal{D}_p\cup\mathcal{D}_i^C)} cg_i \,\mathrm{d}\mu\right] \\
&= \sum_{i\leqslant n} p_i \int_{\mathcal{D}_p\setminus\mathcal{D}} cg_i \,\mathrm{d}\mu - \sum_{i\leqslant n} p_i \int_{\mathcal{D}\setminus\mathcal{D}_p} cg_i \,\mathrm{d}\mu \\
&= c\left(\int_{\mathcal{D}_p\setminus\mathcal{D}} p.g \,\mathrm{d}\mu - \int_{\mathcal{D}\setminus\mathcal{D}_p} p.g \,\mathrm{d}\mu\right)
\end{aligned}$$

And for all $g \in \mathcal{D}_p \setminus \mathcal{D}$, we have $p.g \geqslant 0$ with equality only on the boundary of $\mathcal{D}_p$. So unless $\mathcal{D}_p - \mathcal{D}$ is a set of measure zero, we have

$$\int_{\mathcal{D}_p\setminus\mathcal{D}} p.g \,\mathrm{d}\mu > 0$$

Similarly, for all $g \in \mathcal{D} \setminus \mathcal{D}_p$, we have $p.g < 0$. So unless $\mathcal{D} \setminus \mathcal{D}_p$ is a set of measure zero, we have

$$-\int_{\mathcal{D}\setminus\mathcal{D}_p} p.g \,\mathrm{d}\mu > 0$$

This establishes that

$$\sum_{i\leqslant n} p_i \mathcal{I}_i(\mathcal{D}_p) < \sum_{i\leqslant n} p_i \mathcal{I}_i(\mathcal{D})$$

unless both $\mathcal{D} - \mathcal{D}_p$ and $\mathcal{D}_p - \mathcal{D}$ are sets of measure zero. ■

In general, strictly proper scoring rules generated in this way will not take the additive form of the Brier, Log or Spherical scores.

Example 1 *Let $\Omega = \{\omega_1, \omega_2, \omega_3\}$ and let $\mathcal{P}$ be the set of all probability mass functions of Ω. Choose $p = \langle p_1, p_2, p_3\rangle \in \mathcal{P}$. Let ρ be the normal distribution on the Borel σ-algebra $\mathfrak{B}(\mathbb{R})$ with mean 0 and standard deviation 5. Let μ be the product measure $\rho \times \rho \times \rho$ on $\mathfrak{B}(\mathbb{R}^3)$. In that case*

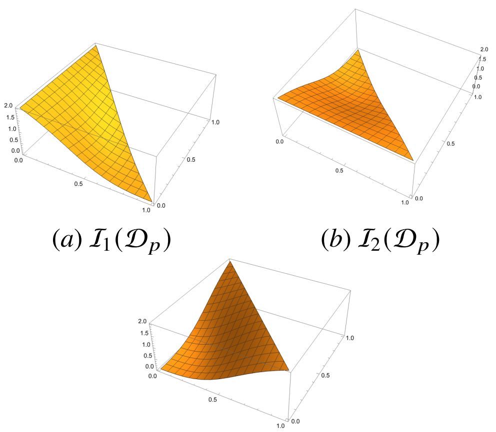
$$\mathcal{I}_i(\mathcal{D}_p) = \frac{5\left(1 - \frac{p_i}{\sqrt{p_1^2+p_2^2+p_3^2}}\right)}{2\pi}$$

See figure 1 for a plot of $\mathcal{I}$. As we will see in example 2, this is a non-additive analogue of the Spherical score. Let $H(p, q) = p.\left\langle \mathcal{I}_1(\mathcal{D}_q), \mathcal{I}_2(\mathcal{D}_q), \mathcal{I}_3(\mathcal{D}_q)\right\rangle$. It is easy to check that $\frac{\partial H}{\partial q_1}(p, q) = \frac{\partial H}{\partial q_2}(p, q) = \frac{\partial H}{\partial q_3}(p, q) = 0$ exactly when $p = q$. Hence $\mathcal{I}$ is strictly proper. But clearly $\mathcal{I}$ does not take the additive form of the Brier, Log or Spherical scores.

Non-additive scoring rules of the sort detailed in example 1 are interesting in part because they provide loss functions for full linear previsions—previsions for an infinite set of gambles. Proposition 2 shows any probability mass function p expects the set of almost desirable gambles determined by p, $\mathcal{D}_p$, to incur lower loss than any arbitrary measurably distinct $\mathcal{D} \subseteq \mathbb{R}^n$, including $\mathcal{D}$ determined by incoherent previsions. Such scoring rules may provide a useful alternative to the strictly proper additive scoring rules for linear previsions considered by [8].

We can recover all additive strictly proper scoring rules as follows. Choose an assignment $c : \mathcal{F} \to \mathbb{R}$ of precise forecasts to events in $\mathcal{F}$. For any $X \in \mathcal{F}$ let

$$\mathcal{D}_{c(X)} = \{\langle g_1, g_2\rangle \,|\, c(X)g_1 + (1 - c(X))g_2 \geqslant 0\}$$



(a) $\mathcal{I}_1(\mathcal{D}_p)$ (b) $\mathcal{I}_2(\mathcal{D}_p)$

(c) $\mathcal{I}_3(\mathcal{D}_p)$

Figure 1: Plot of $\mathcal{I}_i(\mathcal{D}_p)$ as a function of p_1 (x-axis) and p_2 (y-axis).

Let

$$s_0(\mathcal{D}_{c(X)}) = \int_{\mathcal{E}_0} |g_1| \, \mathrm{d}\mu$$

and

$$s_1(\mathcal{D}_{c(X)}) = \int_{\mathcal{E}_1} |g_2| \, \mathrm{d}\mu$$

Proposition 2 implies that the pair $\langle s_0, s_1 \rangle$ is a strictly proper scoring rule (for a single forecast of a single event). [13] shows that we can choose μ appropriately to recover any additive strictly proper scoring rule.

Example 2 *Consider the Spherical Score:*

$$\begin{aligned}\mathcal{I}(c,\omega) &= \sum_{X\in\mathcal{F}} \left(1 - \frac{|1 - \mathbb{1}_X(\omega) - c(X)|}{\sqrt{c(X)^2 + (1-c(X))^2}}\right) \\ &= \sum_{X\in\mathcal{F}:\omega\notin X} \left(1 - \frac{|1 - c(X)|}{\sqrt{c(X)^2 + (1-c(X))^2}}\right) \\ &\quad + \sum_{X\in\mathcal{F}:\omega\in X} \left(1 - \frac{|1 - 1 - c(X)|}{\sqrt{c(X)^2 + (1-c(X))^2}}\right)\end{aligned}$$

Let

$$s_0(c(X)) = \left(1 - \frac{|1 - c(X)|}{\sqrt{c(X)^2 + (1-c(X))^2}}\right)$$

and

$$s_1(c(X)) = \left(1 - \frac{|1 - 1 - c(X)|}{\sqrt{c(X)^2 + (1-c(X))^2}}\right)$$

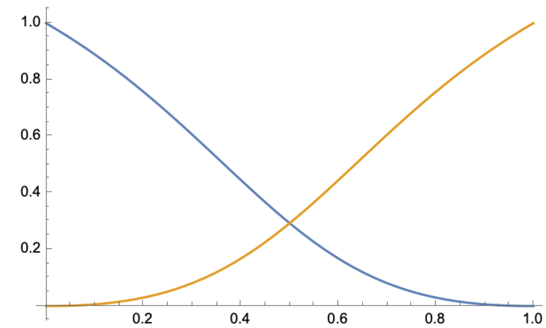


Figure 2: Spherical component scores s_0 (orange) and s_1 (blue).

so that

$$\mathcal{I}(c,\omega) = \sum_{X\in\mathcal{F}:\omega\notin X} s_0(c(X)) + \sum_{X\in\mathcal{F}:\omega\in X} s_1(c(X))$$

We can recover the spherical score by letting ρ be the normal distribution on the Borel σ-algebra $\mathfrak{B}(\mathbb{R})$ with mean 0 and standard deviation 5 and letting μ be the product measure $\rho \times \rho$ on $\mathfrak{B}(\mathbb{R}^2)$. In that case

$$s_0(c(X)) = s_0(\mathcal{D}_{c(X)}) = \int_{\mathcal{E}_0} |g_1| \, \mathrm{d}\mu$$

and

$$s_1(c(X)) = s_1(\mathcal{D}_{c(X)}) = \int_{\mathcal{E}_1} |g_2| \, \mathrm{d}\mu$$

4. Admissibility

We will follow Lindley's basic approach to theorizing about admissibility. Let x be a precise forecast for event E and y be a precise forecast for $\neg E$. The following is a straightforward consequence of [6, Lemma 2]:

Corollary 3 *If $\mathcal{I}_0(x,y) = s_0(x) + s_1(y)$ and $\mathcal{I}_1(x,y) = s_1(x) + s_0(y)$ is a continuously differentiable strictly proper scoring rule, then following three conditions are equivalent:*

1. *There are $a, b \in \mathbb{R}$ s.t.*

$$\nabla_{\langle a,b\rangle}\mathcal{I}_0(x,y) < 0$$

$$\nabla_{\langle a,b\rangle}\mathcal{I}_1(x,y) < 0$$

2. $0 \notin \operatorname{posi}(\{\nabla\mathcal{I}_0(x,y), \nabla\mathcal{I}_1(x,y)\})$

3. $y \neq 1 - x$

So if $y \neq 1 - x$, then there is some point $\langle a, b\rangle$ such that nudging $\langle x, y\rangle$ in the direction of $\langle a, b\rangle$ is guaranteed to decrease inaccuracy (lower your penalty relative to $\mathcal{I}$). Conversely, if $y = 1 - x$, this is never so (since $\mathcal{I}$ is strictly proper). Hence a pair of forecasts, x and y, for E and $\neg E$ respectively, are admissible if and only if probabilistic.

Condition 2 allows us to move between 1 and 3 using a separation theorem. We can proceed in a very similar way when characterising admissible sets of almost desirable gambles relative to our IP scoring rules.

For any linear space $\mathcal{L}$, the **positive hull** of $q_1, \ldots, q_m \in \mathcal{L}$ is

$$\operatorname{posi}(\{q_1, \ldots, q_m\}) = \left\{\sum_{i\leqslant m} a_i q_i \middle| a_1, ..., a_m \geqslant 0, \sum_{i\leqslant m} a_i > 0\right\}$$

The **linear span** of $q_1, \ldots, q_m \in \mathcal{L}$ is

$$\operatorname{span}(\{q_1, \ldots, q_m\}) = \left\{\sum_{i\leqslant m} a_i q_i \middle| a_1, ..., a_m \in \mathbb{R}\right\}$$

Theorem 4 *For any measure space $(X, \mathcal{F}, \nu)$ and any $q_1, \ldots, q_m \in \mathcal{L}^p(X, \mathcal{F}, \nu)$ with $p \geqslant 1$, if*

$$0 \notin posi(\{q_1, \ldots, q_m\})$$

then there is some $h \in \mathcal{L}^{p'}(X, \mathcal{F}, \nu)$, where p' is the conjugate of p, such that for all $i \leqslant m$

$$\int_X q_i h d\nu < 0$$

Proof Suppose $0 \notin \operatorname{posi}(\{q_1, \ldots, q_m\})$. Let

$$A = \operatorname{posi}(\{q_1, \ldots, q_m\}) \cup \{0\}, \qquad B = \{0\}$$

Note that $\mathcal{L}^p(X, \nu)$ with the canonical norm induced topology is a metric space and hence Hausdorff. So A is closed [1, 5.25 Corollary]. Moreover, since $\operatorname{span}(\{q_1, \ldots, q_m\})$ is finite dimensional, it is locally compact, by Riesz's theorem. And every closed subspace of a locally compact space is locally compact. So A is locally compact. And B is trivially closed.

Let $A' = A \cap -A$. Since $0 \notin \operatorname{posi}(\{q_1, \ldots, q_m\})$, $A' = B = \{0\}$.

The previous remarks show that the conditions for Klee's separation theorem [5, Theorem 2.5] hold:

- A and B are closed convex cones in a convex linear space (*i.e.* $\mathcal{L}^p(X, \nu)$)
- A and B have vertex 0
- A is locally compact
- $A' = B = \{0\}$

Klee's separation theorem guarantees that there is a continuous linear functional ψ on $\mathcal{L}^p(X, \nu)$ such that $\phi(a) < 0$ for all $a \in A - A' = \operatorname{posi}(\{q_1, \ldots, q_m\})$ [5, Theorem 2.5]. In particular then $\psi(q_i) < 0$ for all $i \leqslant m$. Finally, by the Riesz-Frechet representation theorem, there is some $h \in \mathcal{L}^{p'}(X, \nu)$, where p' is the conjugate of p, such that for all $b \in \mathcal{L}^p(X, \nu)$

$$\psi(x) = \int_X bh d\nu$$

Hence for all $i \leqslant m$

$$\psi(q_i) = \int_X q_i h d\nu < 0$$

■

We can use theorem 4 to extend corollary 3 to the imprecise case.

Corollary 5 *If $\mathcal{I}$ satisfies the assumptions in section 1 (on $\nu_1, \ldots \nu_n$), then the following two conditions are equivalent:*

1. *There is some $h : \mathbb{R}^{n-1} \to \mathbb{R}$ s.t. for all $i \leqslant n$*

$$\delta\mathcal{I}_i(f, h) < 0$$

2. $0 \notin \operatorname{posi}(\{\phi_i(\cdot, f(\cdot)) \mid i \leqslant n\})$

For notational convenience, let $\mathcal{I}_i(\mathcal{D}_f) = \mathcal{I}_i(f)$.

Proof For any $h : \mathbb{R}^{n-1} \to \mathbb{R}$, the first variation of $\mathcal{I}_i(f)$ is given by

$$\delta\mathcal{I}_i(f, h) = \int_{\mathbb{R}^{n-1}} \frac{\delta\mathcal{I}_i}{\delta f} h \, \mathrm{d}\mu = \int_{\mathbb{R}^{n-1}} \phi_i(\cdot, f(\cdot)) h \, \mathrm{d}\mu$$

Since ν_i is finite for all $i \leqslant n$, $\int_{\mathbb{R}^{n-1}} \phi_i(\cdot, f(\cdot)) \, \mathrm{d}\mu < \infty$. So $\phi_1(\cdot, f(\cdot)), \ldots, \phi_n(\cdot, f(\cdot)) \in \mathcal{L}^1(\mathbb{R}^{n-1})$. Hence by theorem 4, condition 2 implies 1.

To see that 1 implies 2, suppose that there is some $h : \mathbb{R}^{n-1} \to \mathbb{R}$ s.t. for all $i \leqslant n$

$$\delta\mathcal{I}_i(f, h) < 0$$

but $0 = \sum_{i\leqslant m} a_i \phi_i(\cdot, f(\cdot))$ for some $a_1, ..., a_m \geqslant 0$ with $\sum_{i\leqslant m} a_i > 0$. Then

$$\sum_{i\leqslant m} a_i \delta\mathcal{I}_i(f, h) = \sum_{i\leqslant m} a_i \int_{\mathbb{R}^{n-1}} \phi_i(\cdot, f(\cdot)) h \, \mathrm{d}\mu = \int_{\mathbb{R}^{n-1}} \left(\sum_{i\leqslant m} a_i \phi_i(\cdot, f(\cdot))\right) h \, \mathrm{d}\mu < 0,$$

which is a contradiction. ■

Corollary 5 is critical for teasing out axioms on ν_i and ϕ_i from our coherence axioms AD1-AD5. If we specify constraints on ϕ_i which guarantee that $0 \in \operatorname{posi}(\{\phi_i(\cdot, f(\cdot)) \mid i \leqslant n\})$ implies that f satisfies E1-E4, that is sufficient to show that incoherent sets of almost desirable gambles are inadmissible.

To illustrate, consider the following example.

Example 3 *Suppose that $\Omega = \{\omega_1, \omega_2, \omega_3\}$ and that for some $\lambda \geqslant \gamma > 0$:*

$$\phi_i(g_1, g_2, g_3) = \begin{cases} \lambda g_i & \text{if } g_i < 0, \\ \gamma g_i & \text{if } g_i \geqslant 0. \end{cases}$$

Then $0 \in \operatorname{posi}(\{\phi_i(\cdot, f(\cdot)) \mid i \leqslant 3\})$ iff there are $\alpha, \beta > 0$ s.t.

$$f(g_1, g_2) = \begin{cases} \frac{-\gamma(\alpha g_1 + \beta g_2)}{\lambda} & \text{if } g_1 \geqslant 0, g_2 \geqslant 0, \\ \frac{-\lambda(\alpha g_1 + \beta g_2)}{\gamma} & \text{if } g_1 < 0, g_2 < 0, \\ \frac{-(\alpha\lambda g_1 + \beta\gamma g_2)}{\gamma} & \text{if } g_1 < 0, g_2 \geqslant 0, \alpha\lambda g_1 + \beta\gamma g_2 < 0, \\ \frac{-(\alpha\lambda g_1 + \beta\gamma g_2)}{\lambda} & \text{if } g_1 < 0, g_2 \geqslant 0, \alpha\lambda g_1 + \beta\gamma g_2 \geqslant 0, \\ \frac{-(\alpha\gamma g_1 + \beta\lambda g_2)}{\gamma} & \text{if } g_1 \geqslant 0, g_2 < 0, \alpha\gamma g_1 + \beta\lambda g_2 < 0, \\ \frac{-(\alpha\gamma g_1 + \beta\lambda g_2)}{\lambda} & \text{otherwise}. \end{cases}$$

It is easy to verify that any such f satisfies E1-E4 and hence that $\mathcal{D}_f$ is coherent. By corollary 5, then incoherent sets of almost desirable gambles are inadmissible. For any incoherent epigraphical set of almost desirable gambles, $\mathcal{D}_b$, there is some $h : \mathbb{R}^{n-1} \to \mathbb{R}$ such that $\mathcal{D}_{b+\epsilon h}$ is guaranteed to have lower inaccuracy (lower penalty relative to $\mathcal{I}$) for sufficiently small ϵ.

It is certainly possible, then, to specify constraints on ϕ_i that render incoherent sets of almost desirable gambles inadmissible. But we would also like our constraints to be flexible enough so that for any coherent set of almost desirable gambles, there is some $\mathcal{I}$ relative to which it is admissible. Finding the right balance between these two tasks—rendering incoherence admissible while retaining sufficient flexibility in the class of reasonable IP scoring rules—is challenging. We turn to this task in section 5.

5. A Constructive Method for Generating Penalty Functions

Let $\Omega = \{\omega_1, \omega_2, \omega_3\}$ and let $\mathcal{P}$ be the set of all probability mass functions of Ω.

Let $\mathbb{R}_{\geqslant 0} = \{g : g_1, g_2, g_3 \geqslant 0\}$; $\mathbb{R}_{\leqslant 0} = \{g : g_1, g_2, g_3 \leqslant 0\}$.

For any $g \in \mathbb{R}^3 \setminus (\mathbb{R}_{\geqslant 0} \cup \mathbb{R}_{\leqslant 0})$, let

$$L_g = \{p : p.g = 0,\ p_1 + p_2 + p_3 = 1\}$$

Let P be a closed convex set of probability mass functions on Ω such that $\operatorname{int} P \neq \varnothing$ and $P \cap \operatorname{int} \mathbb{P} \neq \varnothing$. Choose $m = \langle m_1, m_2, m_3 \rangle$ in the interior of P. Let $\underline{P}(g) = \inf\{p.g | p \in P\}$ and $\mathcal{D} = \{g | \underline{P}(g) \geqslant 0\}$.

The aim now is to construct a function $\phi : \mathbb{R}^n \to \mathbb{R}^n$ with

$$\phi(g) = \langle \phi_1(g), \phi_2(g), \phi_3(g) \rangle$$

such that $\mathcal{D}$ is admissible relative to the IP scoring rule determined by ϕ, *i.e.*

$$\mathcal{I}_i(\mathcal{D}) = \int_{\mathcal{E}_i} |\phi_i| \, \mathrm{d}\mu$$

and all other admissible sets of almost desirable gambles relative to $\mathcal{I}$ are coherent.

It will be useful in what follows to define a few terms. Choose any $g = \langle g_1, g_2, g_3 \rangle \in \mathbb{R}^3 \setminus (\mathbb{R}_{\geqslant 0} \cup \mathbb{R}_{\leqslant 0})$ with distinct components. Let $\alpha_g, \beta_g, \kappa_g$ be defined as follows:

$$\alpha_g = \left\langle \frac{g_2}{g_2 - g_1}, \frac{-g_1}{g_2 - g_1}, 0 \right\rangle \tag{1}$$

$$\beta_g = \left\langle \frac{g_3}{g_3 - g_1}, 0, \frac{-g_1}{g_3 - g_1} \right\rangle \tag{2}$$

$$\kappa_g = \left\langle 0, \frac{g_3}{g_3 - g_2}, \frac{-g_2}{g_3 - g_2} \right\rangle \tag{3}$$

Note:

1. If g_1, g_2, g_3 are all non-zero, then two of $\alpha_g, \beta_g, \kappa_g$ are probability mass functions and one is not. Moreover, α_g is the unique non-pmf iff g_1 and g_2 have the same sign. β_g is the unique unique non-pmf iff g_1 and g_3 have the same sign. κ_g is the unique unique non-pmf iff g_2 and g_3 have the same sign.

2. If one of g_1, g_2, g_3 is zero, then two of $\alpha_g, \beta_g, \kappa_g$ are identical indicator functions and one is a distinct probability mass function. Moreover, $\alpha_g = \beta_g = \langle 1, 0, 0 \rangle$ iff $g_1 = 0$; $\alpha_g = \kappa_g = \langle 0, 1, 0 \rangle$ iff $g_2 = 0$; $\beta_g = \kappa_g = \langle 0, 0, 1 \rangle$ iff $g_3 = 0$.

Let π_g be the unique element of $\{\alpha_g, \beta_g, \kappa_g\}$ that is not a probability mass function (case 1) or the unique indicator function in $\{\alpha_g, \beta_g, \kappa_g\}$ (case 2).

For any $g = \langle g_1, g_2, g_3 \rangle \in \mathbb{R}^3 \setminus (\mathbb{R}_{\geqslant 0} \cup \mathbb{R}_{\leqslant 0})$ with $\underline{P}(g) = 0$ define $\phi(g)$ as follows. If g_1, g_2, g_3 are all distinct, then $\phi(g)$ is the unique gamble s.t.

$$\pi_g.\phi(g) = m.\phi(g) = 0$$

$\|g\| = \|\phi(g)\|$ and $\operatorname{sign}(g) = \operatorname{sign}(\phi(g))$. If g_1, g_2, g_3 are not all distinct, then $\phi(g)$ is the unique gamble s.t. $m \in L_{\phi(g)}$, $L_g \cap L_{\phi(g)} = \emptyset$, $\|g\| = \|\phi(g)\|$ and $\operatorname{sign}(g) = \operatorname{sign}(\phi(g))$.

Before continuing with the construction of ϕ, it is worth pausing to observe that however we extend ϕ, $\mathcal{D}$ will be admissible relative to $\mathcal{I}$. To see this, note that since $\mathcal{D}$ is coherent, it is epigraphical. Let $\mathcal{D} = \mathcal{D}_f$ so that $\mathcal{D}$ is the epigraph of f. Then for any $g = \langle g_1, g_2, g_3\rangle \in \mathbb{R}^3 \setminus (\mathbb{R}_{\geqslant 0} \cup \mathbb{R}_{\leqslant 0})$ with $\underline{P}(g) = 0$,

$$\phi(g) = \langle \phi_1(g_1, g_2, f(g_1, g_2)), \\ \phi_2(g_1, g_2, f(g_1, g_2)), \phi_3(g_1, g_2, f(g_1, g_2))\rangle$$

And by construction $m.\phi(g) = 0$ for all such g. Hence $0 \in \operatorname{posi}(\{\phi_i(\cdot, f(\cdot)) \mid i \leqslant 3\})$. So by corollary 5 it is admissible.

Return now to our construction. For any $g = \langle g_1, g_2, g_3\rangle \in \mathbb{R}^3 \setminus (\mathbb{R}_{\geqslant 0} \cup \mathbb{R}_{\leqslant 0})$ with $\underline{P}(g) \neq 0$, define g^* and $\phi(g)$ as follows.

If g_1, g_2, g_3 are all distinct, then let g^* be the unique gamble s.t.

$$\pi_g . g^* = \underline{P}(g^*) = 0$$

$\|g\| = \|g^*\|$ and $\operatorname{sign}(g) = \operatorname{sign}(g^*)$. If g_1, g_2, g_3 are not all distinct, then g^* is the unique gamble s.t. $\underline{P}(g^*) = 0$, $L_g \cap L_{g^*} = \emptyset$, $\|g\| = \|g^*\|$ and $\operatorname{sign}(g) = \operatorname{sign}(g^*)$.

Let v_g be defined as follows.

$$v_g = \begin{cases} \mathbb{1}_{\{\omega_1\}} \text{ if } g_1 \geqslant g_2 \geqslant 0, g_3 < 0, \underline{P}(g) < 0 \\ \mathbb{1}_{\{\omega_2\}} \text{ if } g_2 > g_1 \geqslant 0, g_3 < 0, \underline{P}(g) < 0 \\ \mathbb{1}_{\{\omega_3\}} \text{ if } g_1, g_2 \geqslant 0, g_3 < 0, \underline{P}(g) > 0 \\ \mathbb{1}_{\{\omega_1\}} \text{ if } g_1 \leqslant g_2 \leqslant 0, g_3 > 0, \underline{P}(g) > 0 \\ \mathbb{1}_{\{\omega_2\}} \text{ if } g_2 \leqslant g_1 \leqslant 0, g_3 > 0, \underline{P}(g) > 0 \\ \mathbb{1}_{\{\omega_3\}} \text{ if } g_1, g_2 \leqslant 0, g_3 > 0, \underline{P}(g) < 0 \end{cases}$$

Define v_g similarly if (i) $g_1, g_3 \geqslant 0$ and $g_2 < 0$, (ii) $g_1, g_3 \leqslant 0$ and $g_2 > 0$, (iii) $g_2, g_3 \geqslant 0$ and $g_1 < 0$, or (iv) $g_2, g_3 \leqslant 0$ and $g_1 > 0$.

Let δ_g be defined as follows.

$$\delta_g = \begin{cases} \beta_g & \text{if } g_1 \geqslant g_2 \geqslant 0 \text{ and } g_3 < 0 \\ & \text{or } g_1 \leqslant g_2 \leqslant 0 \text{ and } g_3 > 0 \\ \kappa_g & \text{if } g_2 > g_1 \geqslant 0 \text{ and } g_3 < 0 \\ & \text{or } g_2 < g_1 \leqslant 0 \text{ and } g_3 > 0 \\ \alpha_g & \text{if } g_1 \geqslant g_3 \geqslant 0 \text{ and } g_2 < 0 \\ & \text{or } g_1 \leqslant g_3 \leqslant 0 \text{ and } g_2 > 0 \\ \kappa_g & \text{if } g_3 > g_1 \geqslant 0 \text{ and } g_2 < 0 \\ & \text{or } g_3 < g_1 \leqslant 0 \text{ and } g_2 > 0 \\ \alpha_g & \text{if } g_2 \geqslant g_3 \geqslant 0 \text{ and } g_1 < 0 \\ & \text{or } g_2 \leqslant g_3 \leqslant 0 \text{ and } g_1 > 0 \\ \beta_g & \text{if } g_3 > g_2 \geqslant 0 \text{ and } g_1 < 0 \\ & \text{or } g_3 < g_2 \leqslant 0 \text{ and } g_1 > 0 \end{cases}$$

Let $\mu \in (0, 1)$ be the unique value s.t.

$$\delta_g = \mu v_g + (1-\mu)\delta_{g^*}$$

Finally $\phi(g)$ is the unique gamble s.t.

$$\delta_{\phi(g)} = \mu v_g + (1-\mu)\delta_{\phi(g^*)}$$

$\|g\| = \|\phi(g)\|$ and $\operatorname{sign}(g) = \operatorname{sign}(\phi(g))$.

We conjecture that the following two propositions are true.

Conjecture 6 *For any $p \in \mathcal{P}$, $g \in \mathbb{R}^3 \setminus (\mathbb{R}_{\geqslant 0} \cup \mathbb{R}_{\leqslant 0})$ and $\lambda > 0$, if $p.\phi(g) = 0$ then $p.\phi(\lambda g) = 0$.*

Conjecture 7 *For any $p \in \mathcal{P}$, $f, g \in \mathbb{R}^3 \setminus (\mathbb{R}_{\geqslant 0} \cup \mathbb{R}_{\leqslant 0})$ and $0 < \lambda < 1$, if $p.\phi(f) = p.\phi(g) = 0$ then $p.\phi(\lambda f + (1-\lambda)g) \geqslant 0$.*

This construction is lamentably baroque. But it does (i) render our original coherent set of almost desirable gambles admissible, (ii) satisfy P1 and P2, and if conjectures 6 and 7 are correct, then (iii) it renders incoherent sets of almost desirable gambles inadmissible. If successful, it may provide clues toward a fully adequate axiomatization of IP scoring rules.

6. Discussion

The theory of IP scoring rules is still in its infancy. There are many open questions yet to tackle. If the construction in section 5 is successful, we must still generalise it to arbitrary finite dimensional sample spaces and fully characterize reasonable IP scoring rules.

Let us finish with a discussion of what *not* to hope for out of IP scoring rules. Seidenfeld *et al.* [14] establish that there is no strictly proper, continuous real-valued scoring rule for lower and upper probability forecasts. Some lower and upper probability forecasts will be weakly dominated according to any continuous real-valued scoring rule. As a result:

> When the interval $[p, q]$ is the forecaster's IP-uncertainty for event E, she/he will not have reason to announce that interval as her/his forecast rather than the rival forecast $[p', q']$. [14, p. 1256]

Seidenfeld *et al.* abandon the search for continuous real-valued IP scoring rules and opt for a lexicographic scoring rule, which *is* strictly proper for lower/upper probability forecasts of a single event. They are undoubtedly right to be cautious about the limitations of continuous real-valued loss functions. But it is worth emphasising that for IP models to be useful in many domains, we *need* well-behaved, continuous, even differentiable, real-valued IP loss functions. For example, to train neural net classifiers to produce imprecise classification probabilities using standard Python packages, we must be able to calculate the gradient of our real-valued IP loss functions.

Moreover, depending on the purpose at hand, it may not be necessary or even desirable to provide a forecaster with reason to announce her true IP model (lower and upper probabilities, set of almost desirable gambles, etc.). As William James forcefully argued in his exchange with W.K. Clifford:

> He who says "Better to go without belief forever than believe a lie!" merely shows his own preponderant private horror of becoming a dupe. . . Our errors are surely not such awfully solemn things. In a world where we are so certain to incur them in spite of all our caution, a certain lightness of heart seems healthier than this excessive nervousness on their behalf. [3, section VII]

A forecaster may, a bit like Clifford, manifest an "excessive nervousness" over committing type 1 errors (making false judgments of desirability) and as a result have overly imprecise opinions. In that case, it might be a good thing if our IP scoring rule incentivises them to announce a slightly more informative set of almost desirable gambles. That might allow us to extract useful information that could help guide decisions. (Compare: Tetlock's team won the IARPA forecasting tournament by extremizing ensemble forecasts.)

Alternatively, a forecaster may have a pathologically light heart. They may care too much about avoiding type 2 errors (missing out on true judgments of desirability) and not enough about type 1 errors. In that case, it might be a good thing if our IP scoring rule incentivises them to announce a slightly less informative set of almost desirable gambles. For example, public health officials might reasonably want experts to announce IP forecasts *as if* they were more concerned with type 1 errors than they actually are.

Continuous, real-valued IP scoring rules are tools for evaluating IP forecasts on the basis of type 1 and type 2 error. They do, as Seidenfeld *et al.* [14] show, incentivise tactical forecasting. But not all forms of tactical forecasting are problematic. Whether, when and how they are problematic depends on the type of tactical forecasting involved and the purposes of the evaluator.

Acknowledgments

This research was produced as part of the Epistemic Utility for Imprecise Probability project which is funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 852677).

References

[1] Charalambos D. Aliprantis and Kim C. Border. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer, 1999.

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

[3] William James. The will to believe. In *The Will to Believe and Other Essays in Popular Philosophy*. Longmans Green and Co., 1897.

[4] James M Joyce. Accuracy and coherence: Prospects for an alethic epistemology of partial belief. In Franz Huber and Christoph Schmidt-Petri, editors, *Degrees of Belief*, volume 342. Springer, Dordrecht, 2009.

[5] V. L. Klee. Separation properties of convex cones. *Proceedings of the American Mathematical Society*, 6 (2):313–318, 1955.

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

[7] Joseph B. Kadane Mark J. Schervish, Teddy Seidenfeld. Proper scoring rules, dominated forecasts, and coherence. *Decision Analysis*, 6(4), 2009.

[8] Joseph B. Kadane Mark J. Schervish, Teddy Seidenfeld. Infinite previsions and finitely additive expectations. *arXiv:1308.6761*, 2013.

[9] Richard Pettigrew. *Accuracy and the Laws of Credence*. Oxford University Press, Oxford, 2016.

[10] M. Piccioni. Continuous versions of radon-nikodym derivatives as likelihood ratios. *Systems & Control Letters 2*, pages 369–374, 1983.

[11] Joel Predd, Robert Seiringer, Elliott Lieb, Daniel Osherson, H. Vincent Poor, , and Sanjeev R. Kulkarni. Probabilistic coherence and proper scoring rules. *IEEE Transaction on Information Theory*, 55(10): 4786–4792, 2009.

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

[13] Mark Schervish. A general method for comparing probability assessors. *The Annals of Statistics*, 17(4): 1856–1879, 1989.

[14] Teddy Seidenfeld, M. J. Schervish, and J. B. Kadane. Forecasting with imprecise probabilities. *International Journal of Approximate Reasoning*, 53(8):1248–1261., 2012.