★

Objective Bayesianism and the Maximum Entropy Principle

★

Jürgen Landes and Jon Williamson

Draft of September 3, 2013
For Maximum Entropy and Bayes Theorem, a special issue of *Entropy* journal.

Abstract

Objective Bayesian epistemology invokes three norms: the strengths of our beliefs should be probabilities, they should be calibrated to our evidence of physical probabilities, and they should otherwise equivocate sufficiently between the basic propositions that we can express. The three norms are sometimes explicated by appealing to the maximum entropy principle, which says that a belief function should be a probability function, from all those that are calibrated to evidence, that has maximum entropy. However, the three norms of objective Bayesianism are usually justified in different ways. In this paper we show that the three norms can all be subsumed under a single justification in terms of minimising worst-case expected loss. This, in turn, is equivalent to maximising a generalised notion of entropy. We suggest that requiring language invariance, in addition to minimising worst-case expected loss, motivates maximisation of standard entropy as opposed to maximisation of other instances of generalised entropy.

Our argument also provides a qualified justification for updating degrees of belief by Bayesian conditionalisation. However, conditional probabilities play a less central part in the objective Bayesian account than they do under the subjective view of Bayesianism, leading to a reduced role for Bayes' Theorem.

Contents

§1
Introduction

Objective Bayesian epistemology is a theory about strength of belief. As formulated by Williamson (2010), it invokes three norms:

Probability. The strengths of an agent's beliefs should satisfy the axioms of probability. That is, there should be a probability function $P_E : S\mathscr{L} \longrightarrow [0,1]$ such that for each sentence θ of the agent's language $\mathscr{L}$, $P_E(\theta)$ measures the degree to which the agent, with evidence E, believes sentence θ.[1]

Calibration. The strengths of an agent's beliefs should satisfy constraints imposed by her evidence E. In particular, if the evidence determines just that physical probability (aka *chance*) P^* is in some set $\mathbb{P}^*$ of probability functions defined on $S\mathscr{L}$, then P_E should be calibrated to physical probability insofar as it should lie in the convex hull $\mathbb{E} = \langle\mathbb{P}^*\rangle$ of the set $\mathbb{P}^*$.[2]

Equivocation. The agent should not adopt beliefs that are more extreme than is demanded by her evidence E. That is, P_E should be a member of $\mathbb{E}$ that is sufficiently close to the *equivocator* function $P_=$ which gives the same probability to each $\omega \in \Omega$, where the *state descriptions* or *states* ω are sentences describing the most fine-grained possibilities expressible in the agent's language.

One way of explicating these norms proceeds as follows. Measure closeness of P_E to the equivocator by Kullback-Leibler divergence $d(P_E, P_=) = \sum_{\omega\in\Omega} P_E(\omega) \log \frac{P_E(\omega)}{P_=(\omega)}$. Then, if there is some function in $\mathbb{E}$ that is closest to the equivocator, P_E should be such a function. If $\mathbb{E}$ is closed then there is guaranteed to be some function in $\mathbb{E}$ closest to the equivocator; as $\mathbb{E}$ is convex, there is at most one such function. Then we have the maximum entropy principle (Jaynes, 1957): P_E is the function in $\mathbb{E}$ that has maximum entropy H, where $H(P) = -\sum_{\omega\in\Omega} P(\omega)\log P(\omega)$.

The question arises as to how the three norms of objective Bayesianism should be justified, and whether the maximum entropy principle provides a satisfactory explication of the norms.

The Probability norm is usually justified by a *Dutch book* argument. Interpret the strength of an agent's belief in θ to be a *betting quotient*, i.e., a number x such that the agent is prepared to bet xS on θ with return S if θ is true, where S is an unknown stake, positive or negative. Then the only way to avoid the possibility that stakes may be chosen so as to force the agent to lose money, whatever the true state of the world, is to ensure that the betting quotients satisfy the axioms of probability (see, e.g., Williamson, 2010, Theorem 3.2).

The Calibration norm may be justified by a different sort of betting argument. If the agent bets repeatedly on sentences with known chance y with some fixed betting quotient x then she is sure to lose money in the long run unless $x = y$ (see, e.g., Williamson, 2010, pp. 40–41). Alternatively: on a single bet with known chance y, the agent's *expected* loss is positive unless her betting quotient $x = y$, where the expectation is determined with respect to the chance function P^* (Williamson, 2010, pp. 41–42). More generally, if evidence E determines that $P^* \in \mathbb{P}^*$ and the

[1] Here $\mathscr{L}$ will be construed as a finite propositional language and $S\mathscr{L}$ as the set of sentences of $\mathscr{L}$, formed by recursively applying the usual connectives.

[2] We assume throughout this paper that chance is probabilistic, i.e., that P^* is a probability function.

agent makes such bets then sure loss / positive expected loss can be forced unless $P_E \in \langle \mathbb{P}^* \rangle$.

The Equivocation norm may be justified by appealing to a third notion of loss. In the absence of any particular information about the loss $L(\omega, P)$ one incurs when one's strengths of beliefs are represented by P and ω turns out to be the true state, one can argue that one should take the loss function L to be logarithmic, $L(\omega, P) = -\log P(\omega)$ (Williamson, 2010, pp. 64–65). Then the probability function P that minimises worst case expected loss, subject to the information that $P^* \in \mathbb{E}$ where $\mathbb{E}$ is closed and convex, is simply the probability function $P \in \mathbb{E}$ closest to the equivocator—equivalently, the probability function in $\mathbb{E}$ that has maximum entropy (Topsøe, 1979; Grünwald and Dawid, 2004).

The advantage of these three lines of justification is that they make use of the rather natural connection between strength of belief and betting. This connection was highlighted by Frank Ramsey:

> all our lives we are in a sense betting. Whenever we go to the station we are betting that a train will really run, and if we had not a sufficient degree of belief in this we should decline the bet and stay at home. (Ramsey, 1926, p. 183.)

The problem is that the three norms are justified in rather different ways. The Probability norm is motivated by avoiding sure loss. The Calibration norm is motivated by avoiding sure long-run loss, or by avoiding positive expected loss. The Equivocation norm is motived by minimising worst-case expected loss. In particular, the loss function appealed to in the justification of the Equivocation norm differs from that invoked by the justifications of the Probability and Calibration norms.

In this paper we seek to rectify this problem. That is, we seek a single justification of the three norms of objective Bayesian epistemology.

The approach we take is to generalise the justification of the Equivocation norm, outlined above, in order to show that only strengths of beliefs that are probabilistic, calibrated and equivocal minimise worst-case expected loss. We shall adopt the following starting point: as discussed above, $\mathbb{E} = \langle \mathbb{P}^* \rangle$ is taken to be convex and non-empty throughout this paper; we shall also assume that the strengths of the agent's beliefs can be measured by non-negative real numbers—an assumption which is rejected by advocates of ***imprecise probability***, a position which we will discuss separately in §5.3. We do not assume throughout that $\mathbb{E}$ is such that it admits some function that has maximum entropy—e.g., that $\mathbb{E}$ is closed—but we will be particularly interested in the case in which $\mathbb{E}$ does contain its entropy maximiser, in order to see whether some version of the maximum entropy principle is justifiable in that case.

In §2, we shall consider the scenario in which the agent's belief function *bel* is defined over propositions, i.e., sets of possible worlds. Using ω to denote a possible world as well as the state of $\mathscr{L}$ that picks out that possible world, we have that *bel* is a function from the power set of a finite set Ω of possible worlds ω to the non-negative real numbers, $bel : \mathscr{P}\Omega \longrightarrow \mathbb{R}_{\geq 0}$. When it comes to justifying the Probability norm, this will give us enough structure to show that degrees of belief should be additive. Then, in §3, we shall consider the richer framework in which the belief function is defined over sentences, i.e., $bel : S\mathscr{L} \longrightarrow \mathbb{R}_{\geq 0}$. This will allow us to go further by showing that different sentences that express the same proposition should be believed to the same extent. In §4 we shall explain how the preceding results can be used to motivate a version of the maximum entropy principle. In §5 we draw out some of the consequences of our results for Bayes' theorem. In particular, conditional probabilities and Bayes' theorem play a less central role under this approach than they do under subjective Bayesianism. Also in §5, we relate our work to the imprecise probability approach, and suggest that the justification of the norms of objective Bayesianism presented here can be reinterpreted in a non-pragmatic way.

The key results of the paper are intended to demonstrate the following points. Theorem 12 (which deals with beliefs defined over propositions) and Theorem 31 (respectively, belief over sentences) show that only a logarithmic loss function satisfies certain desiderata that, we suggest, any default loss function should satisfy. This allows us to focus our attention on logarithmic loss. Theorems 24, 25 (for propositions), and Theorems 35, 36 (for sentences) show that minimising worst-case expected logarithmic loss corresponds to maximising a generalised notion of entropy. Theorem 39 justifies maximising standard entropy, by viewing this maximiser as a limit of generalised entropy maximisers. Theorem 49 demonstrates a level of agreement between updating beliefs by Bayesian conditionalisation and updating by maximising generalised entropy. Theorem 89 shows that the generalised notion of entropy considered in this paper is pitched at precisely the right level of generalisation.

Three appendices to the paper help to shed light on the generalised notion of entropy introduced in this paper. A motivates the notion by offering justifications of generalised entropy that mirror Shannon's original justification of standard entropy. B explores some of the properties of the functions that maximise generalised entropy. C justifies the level of generalisation of entropy to which we appeal.

§2
Belief over propositions

In this section we shall show that if a belief function defined on propositions is to minimise worst-case expected loss, then it should be a probability function, calibrated to physical probability, which maximises a generalised notion of entropy. The argument will proceed in several steps. As a technical convenience, in §2.1 we shall normalise the belief functions under consideration. In §2.2 we introduce the appropriate generalisation of entropy. In §2.3 we argue that, by default, loss should be taken to be logarithmic. Then in §2.4 we introduce scoring rules, which measure expected loss. Finally, in §2.5 we show that worst-case expected loss is minimised just when generalised entropy is maximised.

For the sake of concreteness we will take Ω to be generated by a propositional language $\mathscr{L} = \{A_1,\ldots,A_n\}$ with propositional variables $A_1,\ldots,A_n$. The states ω take the form $\pm A_1 \wedge \cdots \wedge \pm A_n$ where $+A_i$ is just A_i and $-A_i$ is $\neg A_i$. Thus there are 2^n states $\omega \in \Omega = \{\pm A_1 \wedge \cdots \wedge \pm A_n\}$. We can think of each such state as representing a possible world. A proposition (or, in the terminology of the mathematical theory of probability, an 'event') may be thought of as a subset of Ω, and a belief function $bel : \mathscr{P}\Omega \longrightarrow \mathbb{R}_{\geq 0}$ thus assigns a degree of belief to each proposition that can be expressed in the agent's language. For a proposition $F \subseteq \Omega$ we will use $\bar{F}$ to denote $\Omega \setminus F$. $|F|$ denotes the size of proposition $F \subseteq \Omega$, i.e., the number of states under which it is true.

Let Π be the set of partitions of Ω; a partition $\pi \in \Pi$ is a set of mutually exclusive and jointly exhaustive propositions. To control the proliferation of partitions we shall take the empty set $\emptyset$ to be contained only in one partition, namely $\{\Omega, \emptyset\}$.

§2.1. Normalisation

There are finitely many propositions ($\mathscr{P}\Omega$ has 2^{2^n} members), so any particular belief function *bel* takes values in some interval $[0,M] \subseteq \mathbb{R}_{\geq 0}$. It is just a matter of convention as to the scale on which belief is measured, i.e., as to what upper bound M we might consider. For convenience we shall normalise the scale to the unit interval $[0,1]$, so that all belief functions are considered on the same scale.

Definition 1 (*Normalised belief function on propositions*). Let $M = \max_{\pi\in\Pi} \sum_{F\in\pi} bel(F)$. Given a belief function $bel : \mathscr{P}\Omega \longrightarrow \mathbb{R}_{\geq 0}$ that is not zero everywhere, its *normalisation* $B : \mathscr{P}\Omega \longrightarrow [0,1]$ is defined by setting $B(F) = bel(F)/M$ for each $F \subseteq \Omega$. We shall denote the set of normalised belief functions by $\mathbb{B}$, so

$$\mathbb{B} = \{B : \mathscr{P}\Omega \longrightarrow [0,1] : \sum_{F\in\pi} B(F) \leq 1 \text{ for all } \pi \in \Pi \text{ and } \sum_{F\in\pi} B(F) = 1 \text{ for some } \pi\}.$$

Without loss of generality we rule out of consideration the non-normalised belief function that gives zero degree of belief to each proposition; it will become clear in §2.4 that this belief function is of little interest as it can never minimise worst-case expected loss. For purely technical convenience we will often consider the convex hull $\langle\mathbb{B}\rangle$ of $\mathbb{B}$. In which case we rule into consideration certain belief functions that are not normalised, but which are convex combinations of normalised belief functions. Henceforth, then, we shall focus our attention on belief functions in $\mathbb{B}$ and $\langle\mathbb{B}\rangle$.

Note that we do not impose any further restrictions on the agent's belief function—such as additivity, or the requirement that $B(G) \leq B(F)$ whenever $G \subseteq F$, or that the empty proposition $\emptyset$ has belief zero or the sure proposition Ω is assigned belief one. Our aim is to show that belief functions that do not satisfy such conditions will expose the agent to avoidable loss.

For any $B \in \langle\mathbb{B}\rangle$ and every $F \subseteq \Omega$ we have $B(F) + B(\bar{F}) \leq 1$ because $\{F, \bar{F}\}$ is a partition. Indeed,

$$\sum_{F\subseteq\Omega} B(F) = \frac{1}{2} \cdot \Big(\sum_{F\subseteq\Omega} B(F) + \sum_{F\subseteq\Omega} B(\bar{F}) \Big) \leq \frac{1}{2} \cdot |\mathscr{P}\Omega| = 2^{2^n - 1}. \tag{1}$$

Recall that a subset of $\mathbb{R}^N$ is compact, if and only if it is closed and bounded.

Lemma 2 (*Compactness*). *$\mathbb{B}$ and $\langle\mathbb{B}\rangle$ are compact.*

Proof: $\mathbb{B} \subset \mathbb{R}^{|\mathscr{P}\Omega|}$ is bounded, where $\subset$ denotes strict subset inclusion. Now consider a sequence $(B_t)_{t\in\mathbb{N}} \in \mathbb{B}$ which converges to some $B \in \mathbb{R}^{|\mathscr{P}\Omega|}$. Then for all $\pi \in \Pi$ we find $\sum_{F\in\pi} B(F) \leq 1$. Assume that $B \notin \mathbb{B}$. Thus for all $\pi \in \Pi$ we have $\sum_{F\in\pi} B(F) < 1$. But then there has to exist a $t_0 \in \mathbb{N}$ such that for all $t \geq t_0$ and all $\pi \in \Pi$, $\sum_{F\in\pi} B_t(F) < 1$. This contradicts $B_t \in \mathbb{B}$. Thus, $\mathbb{B}$ is closed and hence compact.

$\langle\mathbb{B}\rangle$ is the convex hull of a compact set. Hence, $\langle\mathbb{B}\rangle \subset \mathbb{R}^{|\mathscr{P}\Omega|}$ is closed and bounded and thus compact. □

We will be particularly interested in the subset $\mathbb{P} \subseteq \mathbb{B}$ of belief functions defined by:

$$\mathbb{P} = \{B : \mathscr{P}\Omega \longrightarrow [0,1] : \sum_{F\in\pi} B(F) = 1 \text{ for all } \pi \in \Pi\}.$$

$\mathbb{P}$ is the set of probability functions:

Proposition 3. *$P \in \mathbb{P}$ if and only if $P : \mathscr{P}\Omega \longrightarrow [0,1]$ satisfies the axioms of probability:*

P1: *$P(\Omega) = 1$ and $P(\emptyset) = 0$.*

P2: *If $F \cap G = \emptyset$ then $P(F) + P(G) = P(F \cup G)$.*

Proof: Suppose $P \in \mathbb{P}$. $P(\Omega) = 1$ because $\{\Omega\}$ is a partition. $P(\emptyset) = 0$ because $\{\Omega, \emptyset\}$ is a partition and $P(\Omega) = 1$. If $F, G \subseteq \Omega$ are disjoint then $P(F) + P(G) = P(F \cup G)$ because $\{F, G, \overline{F \cup G}\}$ and $\{F \cup G, \overline{F \cup G}\}$ are both partitions so $P(F) + P(G) = 1 - P(\overline{F \cup G}) = P(F \cup G)$.

On the other hand, suppose P1 and P2 hold. That $\sum_{F \in \pi} P(F) = 1$ can be seen by induction on the size of π. If $|\pi| = 1$ then $\pi = \{\Omega\}$ and $P(\Omega) = 1$ by P1. Suppose then that $\pi = \{F_1, \ldots, F_{k+1}\}$ for $k \geq 1$. Now $\sum_{i=1}^{k-1} P(F_i) + P(F_k \cup F_{k+1}) = 1$ by the induction hypothesis and $P(F_k \cup F_{k+1}) = P(F_k) + P(F_{k+1})$ by P2, so $\sum_{F \in \pi} P(F) = 1$ as required. □

Example 4 (*Contrasting $\mathbb{B}$ with $\mathbb{P}$*). Using (1) we find $\sum_{F \subseteq \Omega} P(F) = \frac{|\mathscr{P}\Omega|}{2} \geq \sum_{F \subseteq \Omega} B(F)$ for all $P \in \mathbb{P}$ and $B \in \mathbb{B}$. For probability functions $P \in \mathbb{P}$ probability is evenly distributed among the propositions of fixed size in the following sense:

$$\begin{aligned}\sum_{\substack{F \subseteq \Omega \\ |F| = t}} P(F) &= \sum_{\omega \in \Omega} P(\omega) \cdot |\{F \subseteq \Omega \,:\, |F| = t \text{ and } \omega \in F\}| \\ &= \sum_{\omega \in \Omega} P(\omega) \binom{|\Omega| - 1}{t - 1} = \binom{|\Omega| - 1}{t - 1},\end{aligned}$$

where $P(\omega)$ abbreviates $P(\{\omega\})$. For $B \in \mathbb{B}$ and $t > \frac{|\Omega|}{2} \geq 2$ we have in general only the following inequality

$$0 \leq \sum_{\substack{F \subseteq \Omega \\ |F| = t}} B(F) \leq |\{F \subseteq \Omega \,:\, |F| = t\}| = \binom{|\Omega|}{t}.$$

For $B_1 \in \mathbb{B}$ defined as $B_1(\omega) = 1$ for some specific ω and $B_1(F) = 0$ for all other $F \subseteq \Omega$ we have that the lower bound is tight. For $B_2 \in \mathbb{B}$ defined as $B_2(F) = 1$ for $|F| = t$ and $B_2(F) = 0$ for all other $F \subseteq \Omega$ the upper bound is tight.

To illustrate the potentially uneven distribution of beliefs for a $B \in \mathbb{B}$, let A_1, A_2 be the propositional variables in $\mathscr{L}$, so Ω contains four elements. Now consider the $B \in \mathbb{B}$ such that $B(\emptyset) = 0$, $B(F) = \frac{1}{100}$ for $|F| = 1$, $B(F) = \frac{1}{2}$ for $|F| = 2$, $B(F) = \frac{99}{100}$ for $|F| = 3$ and $B(\Omega) = 1$. Note, in particular, that there is no $P \in \mathbb{P}$ such that $B(F) \leq P(F)$ for all $F \subseteq \Omega$.

§2.2. Entropy

The entropy of a probability function is standardly defined as:

$$H_\Omega(P) := -\sum_{\omega \in \Omega} P(\omega) \log P(\omega).$$

We shall adopt the usual convention that $-x \log 0 = x \cdot \infty = \infty$ if $x > 0$ and $0 \log 0 = 0$.

We will need to extend the standard notion of entropy to apply to normalised belief functions, not just to probability functions. Note that the standard entropy only takes into account those propositions that are in the partition $\{\{\omega\} : \omega \in \Omega\}$, which partitions Ω into states. This is appropriate when entropy is applied to probability functions because a probability function is determined by its values on the states. But this is not appropriate if entropy is to be applied to belief functions: in that case one cannot simply disregard all those propositions which are not in the partition of Ω into states—one needs to consider propositions in other partitions too. In fact there are a range of entropies of a belief function, according to how much weight is given to each partition π in the entropy sum:

Definition 5 (*g-entropy*). Given a weighting function $g : \Pi \longrightarrow \mathbb{R}_{\geq 0}$, the *generalised entropy* or *g-entropy* of a normalised belief function is defined as

$$H_g(B) := -\sum_{\pi \in \Pi} g(\pi) \sum_{F \in \pi} B(F) \log B(F).$$

The standard entropy H_Ω corresponds to g_Ω-entropy where

$$g_\Omega(\pi) = \left\{ \begin{array}{lll} 1 & : & \pi = \{\{\omega\} : \omega \in \Omega\} \\ 0 & : & \text{otherwise} \end{array} \right. .$$

We can define the *partition entropy* H_Π to be the g_Π-entropy where $g_\Pi(\pi) = 1$ for all $\pi \in \Pi$. Then

$$\begin{aligned} H_\Pi(B) &= -\sum_{\pi \in \Pi} \sum_{F \in \pi} B(F) \log B(F) \\ &= -\sum_{F \subseteq \Omega} par(F) B(F) \log B(F), \end{aligned}$$

where $par(F)$ is the number of partitions in which F occurs. Note that according to our convention, $par(\emptyset) = 1$ and $par(\Omega) = 2$ because Ω occurs in partitions $\{\emptyset, \Omega\}$ and $\{\Omega\}$. Otherwise, $par(F) = b_{|\bar{F}|}$ where $b_k := \sum_{i=1}^{k} 1/i! \sum_{j=0}^{i} (-1)^{i-j} \binom{i}{j} j^k$ is the k'th *Bell number*, i.e., the number of partitions of a set of k elements.

We can define the *proposition entropy* $H_{\mathscr{P}\Omega}$ to be the $g_{\mathscr{P}\Omega}$-entropy where

$$g_{\mathscr{P}\Omega}(\pi) = \left\{ \begin{array}{lll} 1 & : & |\pi| = 2 \\ 0 & : & \text{otherwise} \end{array} \right. .$$

Then,

$$\begin{aligned} H_{\mathscr{P}\Omega}(B) &= -\sum_{\substack{\pi \in \Pi \\ |\pi| = 2}} \sum_{F \in \pi} B(F) \log B(F) \\ &= -\sum_{F \subseteq \Omega} B(F) \log B(F). \end{aligned}$$

In general, we can express $H_g(B)$ in following way, which reverses the order of the summations,

$$H_g(B) = -\sum_{F \subseteq \Omega} \Big(\sum_{\substack{\pi \in \Pi \\ F \in \pi}} g(\pi) \Big) B(F) \log B(F).$$

As noted above, one might reasonably demand of a measure of the entropy of a belief function that each belief should contribute to the entropy sum, i.e., for each $F \subseteq \Omega$, $\sum_{\substack{\pi \in \Pi \\ F \in \pi}} g(\pi) \neq 0$:

Definition 6 (*Inclusive weighting function*). A weighting function $g : \Pi \longrightarrow \mathbb{R}_{\geq 0}$ is *inclusive* if for all $F \subseteq \Omega$ there is some partition π containing F such that $g(\pi) > 0$.

This desideratum rules out the standard entropy in favour of other candidate measures such as the partition entropy and the proposition entropy.

We have seen so far that g-entropy is a natural generalisation of standard entropy from probability functions to belief functions. In §2.5 we shall see that g-entropy is of particular interest because maximising g-entropy corresponds to minimising worst-case expected loss—this is our main reason for introducing the concept. But there is a third reason why g-entropy is of interest. Shannon (1948, §6) provided an axiomatic justification of standard entropy as a measure of the uncertainty encapsulated in a probability function. Interestingly, as we show in Appendix A, Shannon's argument can be adapted to give a justification of our generalised entropy measure. Thus g-entropy can also be thought of as a measure of the uncertainty of a belief function.

In the remainder of this section we will examine some of the properties of g-entropy.

Lemma 7. *The function* $-\log : [0,1] \to [0,\infty]$ *is continuous in the standard topology on* $\mathbb{R}_{\geq 0} \cup \{+\infty\}$.

Proof: To obtain the standard topology on $\mathbb{R}_{\geq 0} \cup \{+\infty\}$, take as open sets infinite unions and finite intersections over the open sets of $\mathbb{R}_{\geq 0}$ and sets of the form $(r,\infty]$ where $r \in \mathbb{R}$. In this topology on $[0,\infty]$, a set $M \subseteq \mathbb{R}_{\geq 0}$ is open if and only if it is open in the standard topology in $\mathbb{R}_{\geq 0}$. Hence, $-\log$ is continuous in this topology on $(0,1]$.

Let $(a_t)_{t\in\mathbb{N}}$ be a sequence in $[0,1]$ with limit 0. For all $\epsilon > 0$ there exists a $T \in \mathbb{N}$ such that $-\log a_t > \frac{1}{\epsilon}$ for all $t > T$. Hence, for all open sets U containing $+\infty$ there exists a K such that $-\log a_m \in U$, if $m > K$. So, $-\log a_t$ converges to $+\infty$. Thus, $\lim_{t\to\infty} -\log a_t = +\infty = -\log \lim_{t\to\infty} a_t$. □

Proposition 8. *g-entropy is non-negative and, for inclusive g, strictly concave on* $\langle\mathbb{B}\rangle$.

Proof: $B(F) \in [0,1]$ for all F so $\log B(F) \leq 0$, and $g(\pi)\sum_{F\in\pi} B(F)\log B(F) \leq 0$. Hence $\sum_{\pi\in\Pi} -g(\pi)\sum_{F\in\pi} B(F)\log B(F) \geq 0$, i.e., g-entropy is non-negative.

Take distinct $B_1, B_2 \in \langle\mathbb{B}\rangle$ and $\lambda \in (0,1)$ and let $B = \lambda B_1 + (1-\lambda)B_2$. Now, $x\log x$ is strictly convex on $[0,1]$, i.e.,

$$B(F)\log B(F) \leq \lambda B_1(F)\log B_1(F) + (1-\lambda)B_2(F)\log B_2(F)$$

with equality just when $B_1(F) = B_2(F)$.

Consider an inclusive weighting function g.

$$\begin{aligned} H_g(\lambda B_1 + (1-\lambda)B_2) &= -\sum_{\pi\in\Pi} g(\pi)\sum_{F\in\pi} B(F)\log B(F) \\ &\geq -\sum_{\pi\in\Pi} g(\pi)\sum_{F\in\pi} (\lambda B_1(F)\log B_1(F) + (1-\lambda)B_2(F)\log B_2(F)) \\ &= \lambda H_g(B_1) + (1-\lambda)H_g(B_2), \end{aligned}$$

with equality iff for all F, $B_1(F) = B_2(F)$, since g is inclusive. But B_1 and B_2 are distinct so equality does not obtain. In other words, g-entropy is strictly concave. □

Corollary 9. *For inclusive g, if g-entropy is maximised by a function $P^\dagger$ in convex $\mathbb{E} \subseteq \mathbb{P}$, it is uniquely maximised by $P^\dagger$ in $\mathbb{E}$.*

Corollary 10. *For inclusive g, g-entropy is uniquely maximised in the closure* $[\mathbb{E}]$ *of* $\mathbb{E}$.

If g is not inclusive, concavity is not strict. For example, if the standard entropy H_Ω is maximised by $B^\dagger$ then it is also maximised by any belief function $C^\dagger$ that agrees with $B^\dagger$ on the states $\omega \in \Omega$.

Note that different g-entropy measures can have different maximisers on a convex subset $\mathbb{E}$ of probability functions. For example, when $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$ and $\mathbb{E} = \{P \in \mathbb{P} : P(\omega_1) + 2.75P(\omega_2) + 7.1P(\omega_3) = 1.7,\ P(\omega_4) = 0\}$ then the proposition entropy maximiser, the standard entropy maximiser and the partition entropy maximiser are all different, as can be seen from Fig. 1.

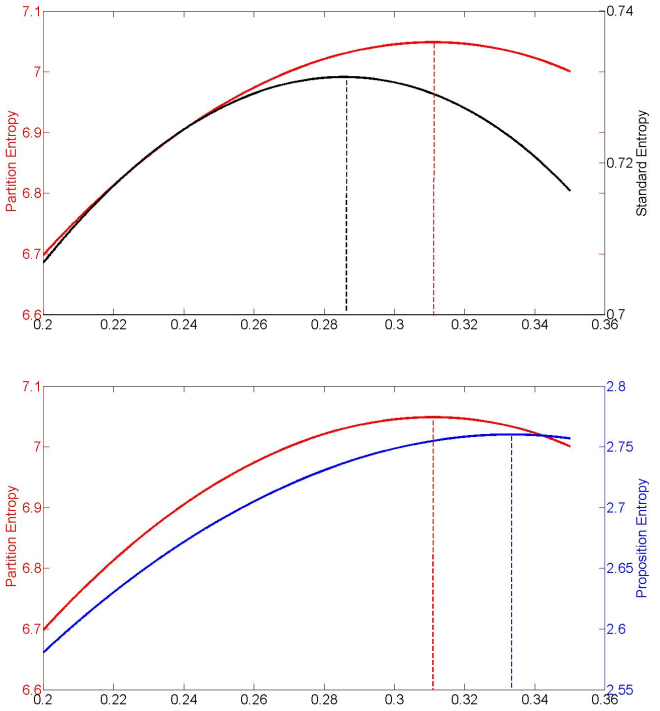

Figure 1: Plotted are the partition entropy, the standard entropy and the proposition entropy under the constraints $P(\omega_1)+P(\omega_2)+P(\omega_3)+P(\omega_4) = 1$, $P(\omega_1)+2.75P(\omega_2)+7.1P(\omega_3) = 1.7, P(\omega_4) = 0$ as a function of $P(\omega_2)$. The dotted lines indicate the respective maxima which obtain for different values of $P(\omega_2)$.

§2.3. Loss

As Ramsey observed, all our lives we are in a sense betting. The strengths of our beliefs guide our actions and expose us to possible losses. If we go to the station when the train happens not to run, we incur a loss: a wasted journey to the station and a delay in getting to where we want to go. Normally, when we are deliberating about how strongly to believe a proposition, we have no realistic idea as to the losses that that belief will expose us to. That is, when determining a belief function B we do not know the true loss function L^*.

Now a loss function L is standardly defined as a function $L : \Omega \times \mathbb{P} \longrightarrow (-\infty,\infty]$, where $L(\omega,P)$ is the loss one incurs by adopting probability function $P \in \mathbb{P}$ when ω is the true state of the world. Note that a standard loss function will only evaluate an agent's beliefs about the states, not the extent to which she believes other propositions. This is appropriate when belief is assumed to be probabilistic, because a probability function is determined by its values on the states. But we are concerned with justifying the Probability norm here and hence need to consider the full range of the agent's beliefs, in order to show that they should satisfy the axioms of probability. Hence we need to extend the concept of a loss function to evaluate all of the agent's beliefs:

Definition 11 (*Loss function*). A *loss function* is a function $L : \mathscr{P}\Omega \times \langle\mathbb{B}\rangle \longrightarrow (-\infty,\infty]$.

$L(F,B)$ is the loss incurred by a belief function B when proposition F turns out to be true. We shall interpret this loss as the loss that is attributable to F in isolation from all other propositions, rather than the *total* loss incurred when proposition F turns out to be true. When F turns out to be true so does any proposition G for $F \subset G$. Thus the total loss when F turns out true includes $L(G,B)$ as well as $L(F,B)$. The total loss on F turning out true might therefore be represented by $\sum_{G \supseteq F} L(G,B)$, with $L(F,B)$ being the loss distinctive to F, i.e., the loss on F turning out true *over and above* the loss incurred by $G \supset F$.

Is there anything that one can presume about a loss function in the absence of any information about the true loss function L^*? Plausibly:[3]

L1. $L(F,B) = 0$ if $B(F) = 1$.

L2. $L(F,B)$ strictly increases as $B(F)$ decreases from 1 towards 0.

L3. $L(F,B)$ depends only on $B(F)$.[4]

To express the next condition we need some notation. Suppose $\mathscr{L} = \mathscr{L}_1 \cup \mathscr{L}_2$: say that $\mathscr{L} = \{A_1,\dots,A_n\}$, $\mathscr{L}_1 = \{A_1,\dots,A_m\}$, $\mathscr{L}_2 = \{A_{m+1},\dots,A_n\}$ for some $1 < m < n$. Then $\omega \in \Omega$ takes the form $\omega_1 \wedge \omega_2$ where $\omega_1 \in \Omega_1$ is a state of $\mathscr{L}_1$, and $\omega_2 \in \Omega_2$ is a state of $\mathscr{L}_2$. Given propositions $F_1 \subseteq \Omega_1$ and $F_2 \subseteq \Omega_2$ we can define $F_1 \times F_2 := \{\omega = \omega_1 \wedge \omega_2 : \omega_1 \in F_1, \omega_2 \in F_2\}$, a proposition of $\mathscr{L}$. Given a fixed belief function B such that $B(\Omega) = 1$, $\mathscr{L}_1$ and $\mathscr{L}_2$ are *independent sublanguages*, written $\mathscr{L}_1 \perp\!\!\!\perp_B \mathscr{L}_2$, if $B(F_1 \times F_2) = B(F_1) \cdot B(F_2)$ for all $F_1 \subseteq \Omega_1$ and $F_2 \subseteq \Omega_2$, where $B(F_1) := B(F_1 \times \Omega_2)$ and $B(F_2) := B(\Omega_1 \times F_2)$. The restriction $B_{\restriction \mathscr{L}_1}$ of B to $\mathscr{L}_1$ is a belief function on $\mathscr{L}_1$ defined by $B_{\restriction \mathscr{L}_1}(F_1) = B(F_1) = B(F_1 \times \Omega_2)$, and similarly for $\mathscr{L}_2$.

L4. Losses are additive when the language is composed of independent sublanguages: if $\mathscr{L} = \mathscr{L}_1 \cup \mathscr{L}_2$ for $\mathscr{L}_1 \perp\!\!\!\perp_B \mathscr{L}_2$ then $L(F_1 \times F_2, B) = L_1(F_1, B_{\restriction \mathscr{L}_1}) + L_2(F_2, B_{\restriction \mathscr{L}_2})$, where L_1, L_2 are loss functions defined on $\mathscr{L}_1, \mathscr{L}_2$ respectively.

[3] These conditions correspond to conditions L1–4 of Williamson (2010, pp. 64–65) which were put forward in the special case of loss functions defined over probability functions as opposed to belief functions.

[4] This condition, which is sometimes called *locality*, rules out that $L(F,B)$ depends on $B(F')$ for $F' \neq F$. It also rules out a dependence on $|F|$, for instance.

L1 says that one should presume that fully believing a true proposition will not incur loss. L2 says that one should presume that the less one believes a true proposition, the more loss will result. L3 expresses the interpretation of $L(F,B)$ as the loss attributable to F in isolation of all other propositions. L4 expresses the intuition that, at least if one supposes two propositions to be unrelated, one should presume that the loss on both turning out true is the sum of the losses on each.

The four conditions taken together tightly constrain the form of a presumed loss function L:

Theorem 12. *If loss functions are assumed to satisfy L1–4 then* $L(F,B) = -k\log B(F)$ *for some constant* $k > 0$ *that does not depend on* $\mathscr{L}$*.*

Proof: We shall first focus on a loss function L defined with respect to a language $\mathscr{L}$ that contains at least two propositional variables.

L3 implies that $L(F,B) = f_{\mathscr{L}}(B(F))$, for some function $f_{\mathscr{L}} : [0,1] \longrightarrow (-\infty,\infty]$.

For our fixed $\mathscr{L}$ and each $x,y \in [0,1]$ choose some particular $B \in \langle\mathbb{B}\rangle, \mathscr{L}_1, \mathscr{L}_2, F_1 \subseteq \Omega_1, F_2 \subseteq \Omega_2$ such that $\mathscr{L} = \mathscr{L}_1 \cup \mathscr{L}_2$ where $\mathscr{L}_1 \perp\!\!\!\perp_B \mathscr{L}_2$, $B(F_1) = x$ and $B(F_2) = y$. This is possible because $\mathscr{L}$ has at least two propositional variables. Note in particular that since $\mathscr{L}_1$ and $\mathscr{L}_2$ are independent sublanguages we have $B(\Omega) = 1$.

Note that
$$1 = B(\Omega) = B(\Omega_1 \times \Omega_2) = B_{\restriction \mathscr{L}_1}(\Omega_1),$$
and similarly $B_{\restriction \mathscr{L}_2}(\Omega_2) = 1$. By L1, then, $L_1(\Omega_1, B_{\restriction \mathscr{L}_1}) = L_2(\Omega_2, B_{\restriction \mathscr{L}_2}) = 0$.

So by applying L4 twice:

$$\begin{aligned}
f_{\mathscr{L}}(xy) &= f_{\mathscr{L}}(B(F_1)\cdot B(F_2)) \\
&= L(F_1 \times F_2, B) \\
&= L_1(F_1, B_{\restriction \mathscr{L}_1}) + L_2(F_2, B_{\restriction \mathscr{L}_2}) \\
&= [L(F_1 \times \Omega_2, B) - L_2(\Omega_2, B_{\restriction \mathscr{L}_2})] + [L(\Omega_1 \times F_2, B) - L_1(\Omega_1, B_{\restriction \mathscr{L}_1})] \\
&= L(F_1 \times \Omega_2, B) + L(\Omega_1 \times F_2, B) \\
&= f_{\mathscr{L}}(x) + f_{\mathscr{L}}(y).
\end{aligned}$$

The negative logarithm on $(0,1]$ is characterisable up to a multiplicative constant $k_{\mathscr{L}}$ in terms of this additivity, together with the condition that $f_{\mathscr{L}}(x) \geq 0$ which is implied by L1–2 (see, e.g., Aczél and Daróczy, 1975, Theorem 0.2.5). L2 ensures that $f_{\mathscr{L}}$ is not zero everywhere, so $k_{\mathscr{L}} > 0$.

We thus know that $f_{\mathscr{L}}(x) = -k_{\mathscr{L}} \log x$ for $x \in (0,1]$. Now note that for all $y \in (0,1]$ it needs to be the case that $f_{\mathscr{L}}(0) = f_{\mathscr{L}}(0 \cdot y) = f_{\mathscr{L}}(0) + f_{\mathscr{L}}(y)$, if $f_{\mathscr{L}}$ is to satisfy $f_{\mathscr{L}}(x \cdot y) = f_{\mathscr{L}}(x) + f_{\mathscr{L}}(y)$ for all $x,y \in [0,1]$. Since $f_{\mathscr{L}}$ takes values in $(-\infty,+\infty]$ it follows that $f_{\mathscr{L}}(0) = +\infty$.

Thus far we have shown that for a fixed language $\mathscr{L}$ with at least two propositional variables, $L(F,B) = -k_{\mathscr{L}} \log B(F)$ on $[0,1]$.

Now consider an arbitrary language $\mathscr{L}_1$ and a loss function L_1 on $\mathscr{L}_1$ which satisfies L1 – L4. There exists some other language $\mathscr{L}_2$ and a belief function B on $\mathscr{L} = \mathscr{L}_1 \cup \mathscr{L}_2$ such that $\mathscr{L}_1 \perp\!\!\!\perp_B \mathscr{L}_2$. By the above, for the loss function L on $\mathscr{L}$ it holds that $L(F,B) = -k_{\mathscr{L}} \log B(F)$ on $[0,1]$. By reasoning analogous to that above,

$$L_1(F_1, B_{\restriction \mathscr{L}_1}) = L(F_1 \times \Omega_2, B) = f_{\mathscr{L}}(B(F_1 \times \Omega_2)) = f_{\mathscr{L}}(B_{\restriction \mathscr{L}_1}(F_1)).$$

So the loss function for $\mathscr{L}_1$ is $L_1(F_1, B_{\restriction \mathscr{L}_1}) = -k_{\mathscr{L}} \log B_{\restriction \mathscr{L}_1}(F_1)$. Thus the constant $k_{\mathscr{L}}$ does not depend on the particular language $\mathscr{L}$ after all.

In general, then, $L(F,B) = -k\log B(F)$ for some positive k. □

Since multiplication by a constant is equivalent to change of base, we can take log to be the natural logarithm. Since we will be interested in the belief functions that minimise loss, rather than in the absolute value of any particular losses, we can take $k = 1$ without loss of generality. Theorem 12 thus allows us to focus on the logarithmic loss function:

$$L^{\log}(F,B) := -\log B(F).$$

§2.4. Score

In this paper we are concerned with showing that the norms of objective Bayesianism must hold if an agent is to control her worst-case expected loss. Now an *expected loss function* or *scoring rule* is standardly defined as $S^L_\Omega : \mathbb{P}\times\mathbb{P} \longrightarrow [-\infty,\infty]$ such that $S^L_\Omega(P,Q) = \sum_{\omega\in\Omega} P(\omega)L_\Omega(\omega,Q)$. This is interpretable as the expected loss incurred by adopting probability function Q as one's belief function, when the probabilities are actually determined by P.[5] While this standard definition of scoring rule is entirely appropriate when belief is assumed to be probabilistic, we make no such assumption here and need to consider scoring rules that evaluate all the agent's beliefs, not just those concerning the states. In line with our discussion of entropy in §2.2, we shall consider the following generalisation:

Definition 13 (*g-score*). Given a loss function L and an inclusive weighting function $g : \Pi \longrightarrow \mathbb{R}_{\geq 0}$, the *g-expected loss function* or *g-scoring rule* or simply *g-score* is $S^L_g : \mathbb{P}\times\langle\mathbb{B}\rangle \longrightarrow [-\infty,\infty]$ such that

$$S^L_g(P,B) = \sum_{\pi\in\Pi} g(\pi)\sum_{F\in\pi} P(F)L(F,B).$$

Clearly S^L_Ω corresponds to $S^L_{g_\Omega}$ where g_Ω, which is not inclusive, is defined as in §2.2. We require that g be inclusive in Definition 13, since only in that case does the g-score genuinely evaluate all the agent's beliefs. We will focus on $S^{\log}_g(P^*,B)$, i.e., the case in which the loss function is logarithmic and the expectation is taken with respect to the chance function P^*, in order to show that an agent should satisfy the norms of objective Bayesianism if she is to control her worst-case g-expected logarithmic loss when her evidence determines that the chance function P^* is in $\mathbb{E}$.

For example, with the logarithmic loss function, the *partition* Π-score is defined by setting $g = g_\Pi$:

$$S^{\log}_\Pi(P,B) = -\sum_{\pi\in\Pi}\sum_{F\in\pi} P(F)\log B(F).$$

[5] This is the standard statistical notion of a scoring rule as defined in Dawid (1986). More recently a different, 'epistemic' notion of scoring rule has been considered in the literature on non-pragmatic justifications of Bayesian norms; see, e.g., Joyce (2009); Pettigrew (2011), and also a forthcoming paper by Landes where similarities and differences of these two notions of a scoring rule are discussed. One difference which is significant to our purposes is that Predd et al.'s result in Predd et al. (2009)—that for every epistemic scoring rule which is continuous and strictly proper, the set of non-dominated belief functions is the set $\mathbb{P}$ of probability functions—does not apply to statistical scoring rules. Also, Predd et al. are only interested in justifying the Probability norm by appealing to dominance as a decision theoretic norm. We are concerned with justifying three norms (all at once) using worst-case loss avoidance as a desideratum. The epistemic approach is considered further in §5.4.

Similarly, the *proposition* $\mathscr{P}\Omega$-score is defined by setting $g = g_{\mathscr{P}\Omega}$:

$$S^{\log}_{\mathscr{P}\Omega}(P,B) = -\sum_{F\subseteq\Omega} P(F)\log B(F).$$

It turns out that the various logarithmic scoring rules have the following useful property:[6]

Definition 14 (*Strictly proper g-score*). A scoring rule $S^L_g : \mathbb{P}\times\langle\mathbb{B}\rangle \longrightarrow [-\infty,\infty]$ is *strictly proper* if for all $P\in\mathbb{P}$, the function $S^L_g(P,\cdot) : \langle\mathbb{B}\rangle \longrightarrow [-\infty,\infty]$ has a unique global minimum at $B = P$.

On the way to showing that logarithmic g-scores are strictly proper, it will be useful to consider the following natural generalisation of Kullback-Leibler divergence to our framework:

Definition 15 (*g-divergence*). For a weighting function $g : \Pi \longrightarrow \mathbb{R}_{\geq 0}$, the g-divergence is the function $d_g : \mathbb{P}\times\langle\mathbb{B}\rangle \longrightarrow [-\infty,\infty]$ defined by

$$d_g(P,B) = \sum_{\pi\in\Pi} g(\pi) \sum_{F\in\pi} P(F)\log\frac{P(F)}{B(F)}.$$

Here we adopt the usual convention that $0\log\frac{0}{0} = 0$ and $x\log\frac{x}{0} = +\infty$ for $x\in(0,1]$.

We shall see that $d_g(P,B)$ is a sensible notion of the divergence of P from B by appealing to the following useful inequality (see, e.g., Cover and Thomas, 1991, Theorem 2.7.1):

Lemma 16 (*Log sum inequality*). *For $x_i, y_i \in \mathbb{R}_{\geq 0}, i,j = 1,\ldots,k$,*

$$(\sum_{i=1}^{n} x_i)\log\frac{\sum_{i=1}^n x_i}{\sum_{i=1}^n y_i} \leq \sum_{i=1}^{n} x_i \log\frac{x_i}{y_i}$$

with equality iff $x_i = cy_i$ for some constant c and $i = 1,\ldots,k$.

Proposition 17. *The following are equivalent:*

- $d_g(P,B) \geq 0$ *with equality iff* $B = P$.
- g *is inclusive.*

[6] Definition 14 can be generalised: a scoring rule is *strictly* $\mathbb{X}$*-proper* if it is strictly proper for belief functions taken to be from a set $\mathbb{X}$. In Definition 14, $\mathbb{X} = \langle\mathbb{B}\rangle$. The logarithmic scoring rule in the standard sense, i.e., $\sum_{\omega\in\Omega} P(\omega)L(\omega,Q)$, is well known to be the only strictly $\mathbb{P}$-proper local scoring rule—see McCarthy (1956, p. 654) who credits Andrew Gleason for the uniqueness result; Shuford et al. (1966, p. 136) for the case of continuous scoring rules; Aczel and Pfanzagl (1967, Theorem 3, p. 101) for the case of differentiable scoring rules; and Savage (1971, §9.4). Logarithmic score in our sense, i.e., $\sum_{F\subseteq\Omega} P(F)L(F,B)$, is not strictly $\mathbb{Y}$-proper when $\mathbb{Y}$ is the set of non-normalised belief functions: $S(P,bel)$ is a global minimum, where bel is the belief function such that $bel(F) = 1$ for all F. (While Joyce (2009, p. 276) suggests that logarithmic score is strictly $\mathbb{Y}$-proper for $\mathbb{Y}$ a set of non-normalised belief functions, he is referring to a logarithmic scoring rule that is different to the usual one considered above and that does not satisfy the locality condition L3.)

Proof: First we shall see that if g is inclusive then $d_g(P,B) \geq 0$ with equality iff $B = P$.

$$\begin{aligned} d_g(P,B) &= \sum_{\pi\in\Pi} g(\pi) \sum_{F\in\pi} P(F) \log \frac{P(F)}{B(F)} \\ &\geq \sum_{\pi\in\Pi} g(\pi) \left[\left(\sum_{F\in\pi} P(F) \right) \log \frac{\sum_{F\in\pi} P(F)}{\sum_{F\in\pi} B(F)} \right] \\ &\geq \sum_{\pi\in\Pi} g(\pi) \left[1 \log \frac{1}{1} \right] \\ &= 0, \end{aligned}$$

where the first inequality is an application of the log-sum inequality and the second inequality is a consequence of B being in $\langle\mathbb{B}\rangle$. There is equality at the first inequality iff for all $F \subseteq \Omega$ and all π such that $F \in \pi$ and $g(\pi) > 0$, $P(F') = c_\pi B(F')$ for all $F' \in \pi$, and equality at the second inequality iff for all π such that $g(\pi) > 0$, $\sum_{F\in\pi} B(F) = 1$.

Clearly if $B(F) = P(F)$ for all F then these two equalities obtain. Conversely, suppose the two equalities obtain. Then for each F there is some $\pi = \{F = F_1, F_2, \ldots, F_k\}$ such that $g(\pi) > 0$, because g is inclusive. The first equality condition implies that $P(F_i) = c_\pi B(F_i)$ for $i = 1, \ldots, k$. The second equality implies that $\sum_{i=1}^k B(F_i) = 1$. Hence, $1 = \sum_{i=1}^k P(F_i) = c_\pi \sum_{i=1}^k B(F_i) = c_\pi$, and so $P(F_i) = B(F_i)$ for $i = 1, \ldots, k$. In particular, $B(F) = P(F)$.

Next we shall see that the condition that g is inclusive is essential.

If g were not inclusive then there would be some $F \subseteq \Omega$ such that $g(\pi) = 0$ for all $\pi \in \Pi$ such that $F \in \pi$. There are two cases.

(i) $\emptyset \subset F \subset \Omega$. Take some $P \in \mathbb{P}$ such that $P(F) > 0$. Now define $B(F) := 0$, and $B(F') := P(F')$ for all other F'. Then $B(\Omega) = 1$ and $\sum_{G\in\pi} B(G) \leq 1$ for all other $\pi \in \Pi$, so $B \in \mathbb{B} \subseteq \langle\mathbb{B}\rangle$. Furthermore, $d_g(P,P) = d_g(P,B) = 0$.

(ii) $F = \emptyset$ or $F = \Omega$. Define $B(\emptyset) := B(\Omega) := 0.5$ and $B(F) := P(F)$ for all $\emptyset \subset F \subset \mathscr{P}\Omega$. Then $B(\emptyset) + B(\Omega) = 1$ and $\sum_{G\in\pi} B(G) \leq 1$ for all other $\pi \in \Pi$, so $B \in \mathbb{B} \subseteq \langle\mathbb{B}\rangle$. Furthermore, $d_g(P,P) = d_g(P,B) = 0$.

In either case, then, $d_g(P,B)$ is not uniquely minimised by $B = P$. □

Corollary 18. *The logarithmic g-score is strictly proper.*

Proof: Recall that in the context of a g-score, g is inclusive.

$$\begin{aligned} S_g^{\log}(P,B) - S_g^{\log}(P,P) &= -\sum_{\pi\in\Pi} g(\pi) \sum_{F\in\pi} P(F) \log \frac{B(F)}{P(F)} \\ &= \sum_{\pi\in\Pi} g(\pi) \sum_{F\in\pi} P(F) \log \frac{P(F)}{B(F)} \\ &= d_g(P,B). \end{aligned}$$

Proposition 17 then implies that $S_g^{\log}(P,B) - S_g^{\log}(P,P) \geq 0$ with equality iff $B = P$, i.e., $S_g^{\log}$ is strictly proper. □

Finally, logarithmic g-scores are non-negative strictly convex functions in the following qualified sense:

Proposition 19. *Logarithmic g-score $S_g^{\log}(P,B)$ is non-negative and convex as a function of $B \in \langle\mathbb{B}\rangle$. Convexity is strict, i.e., $S_g^{\log}(P,\lambda B_1 + (1-\lambda)B_2) < \lambda S_g^{\log}(P,B_1) + (1-\lambda)S_g^{\log}(P,B_2)$ for $\lambda \in (0,1)$, unless B_1 and B_2 agree everywhere except where $P(F) = 0$.*

Proof: Logarithmic g-score is non-negative because $B(F), P(F) \in [0,1]$ for all F so $\log B(F) \le 0$, $P(F)\log B(F) \le 0$, and $g(\pi) > 0$.

That $S_g^{\log}(P,B)$ is strictly convex as a function of $\langle\mathbb{B}\rangle$ follows from the strict concavity of $\log x$. Take distinct $B_1, B_2 \in \langle\mathbb{B}\rangle$ and $\lambda \in (0,1)$ and let $B = \lambda B_1 + (1-\lambda)B_2$. Now,

$$\begin{aligned}
P(F)\log B(F) &= P(F)\log(\lambda \cdot B_1(F) + (1-\lambda)B_2(F)) \\
&\ge P(F)\Big(\lambda \log B_1(F) + (1-\lambda)\log B_2(F)\Big) \\
&= \lambda P(F)\log B_1(F) + (1-\lambda)P(F)\log B_2(F)
\end{aligned}$$

with equality iff either $P(F) = 0$ or $B_1(F) = B_2(F)$.

Hence,

$$\begin{aligned}
S_g^{\log}(P,B) &= -\sum_{\pi\in\Pi} g(\pi) \sum_{F\in\pi} P(F)\log B(F) \\
&\le \lambda S_g^{\log}(P,B_1) + (1-\lambda)S_g^{\log}(P,B_2),
\end{aligned}$$

with equality iff B_1 and B_2 agree everywhere except possibly where $P(F) = 0$. □

§2.5. Minimising worst-case logarithmic g-score

In this section we shall show that the g-entropy maximiser minimises worst-case logarithmic g-score.

In order to prove our main result (Theorem 24) we would like to apply a game-theoretic minimax theorem which will allow us to conclude that

$$\inf_{B\in\mathbb{B}} \sup_{P\in\mathbb{E}} S_g^{\log}(P,B) = \sup_{P\in\mathbb{E}} \inf_{B\in\mathbb{B}} S_g^{\log}(P,B).$$

Note that the expression on the left-hand side describes minimising worst-case g-score, where the worst case refers to P ranging in $\mathbb{E}$. Speaking in game-theoretic lingo: the player playing first on the left-hand side aims to find the belief function(s) which minimises worst-case g-expected loss; again the worst case is taken with respect to varying P.

For this approach to work, we would normally need $\mathbb{B}$ to be some set of mixed strategies. It is not obvious how $\mathbb{B}$ could be represented as a mixing of finitely many pure strategies. However, there exists a broad literature on minimax theorems (Ricceri, 2008) and we shall apply a theorem proved in König (1992). This theorem requires that certain level sets, in the set of functions in which the player aiming to minimise may chose his functions, are connected. To apply König's result we will thus allow the belief functions B to range in $\langle\mathbb{B}\rangle$, which has this property. It will follow that the $B \in \langle\mathbb{B}\rangle \setminus \mathbb{B}$ are never good choices for the minimising player playing first: the best choice is in $\mathbb{E}$ which is a subset of $\mathbb{B}$.

Having established that the inf and the sup commute, the rest is straightforward. Since the scoring rule we employ, $S_g^{\log}$, is strictly proper, we have that the best strategy for the minimising player, answering a move by the maximising player, is to select the same function as the maximising player. Thus, it is best for the maximising player playing first to choose a/the function which maximises $S_g^{\log}(P,P)$. We will thus find that

$$\sup_{P\in\mathbb{E}} \inf_{B\in\langle\mathbb{B}\rangle} S_g^{\log}(P,B) = \sup_{P\in\mathbb{E}} \inf_{B\in\{P\}} S_g^{\log}(P,B) = \sup_{P\in\mathbb{E}} S_g^{\log}(P,P) = \sup_{P\in\mathbb{E}} H_g(P).$$

Thus, worst-case g-expected loss and g-entropy have the same value. In game-theoretic terms: we find that our zero-sum g-log-loss game has a value. It remains to be shown that both players, when playing first, have a unique best choice $P^\dagger$.

First, then, we shall apply König's result.

Definition 20 (*König (1992, p. 56)*). For $F:\mathbb{X}\times\mathbb{Y}\to[-\infty,\infty]$ we call $I\subset\mathbb{R}$ a *border interval* of F, if and only if I is an interval of the form $I=(\sup_{x\in\mathbb{X}}\inf_{y\in\mathbb{Y}}F(x,y),+\infty)$. $\Lambda\subset\mathbb{R}$ is called a *border set* of F, if and only if $\inf\Lambda=\sup_{x\in\mathbb{X}}\inf_{y\in\mathbb{Y}}F(x,y)$.

For $\lambda\in\mathbb{R}$ and $\emptyset\subset K\subseteq\mathbb{Y}$ define s_λ and σ_λ to consist of $\mathbb{X}$ and of subsets of $\mathbb{X}$ of the form

$$\bigcap_{y\in K}[F(\cdot,y)>\lambda] \text{ respectively } \bigcap_{y\in K}[F(\cdot,y)\geq\lambda]\ .$$

For $\lambda\in\mathbb{R}$ and finite $\emptyset\subset H\subseteq\mathbb{X}$ define t_λ and τ_λ to consist of subsets of $\mathbb{Y}$ of the form

$$\bigcap_{x\in H}[F(x,\cdot)<\lambda] \text{ respectively } \bigcap_{x\in H}[F(x,\cdot)\leq\lambda]\ .$$

The following may be found in König (1992, Theorem 1.3, p. 57):

Lemma 21 (*König's Minimax*). *Let $\mathbb{X},\mathbb{Y}$ be topological spaces, $\mathbb{Y}$ be compact and Hausdorff and let $F:\mathbb{X}\times\mathbb{Y}\to[-\infty,\infty]$ be lower semicontinuous. Then, if Λ is some border set and I some border interval of F and if at least one of the following conditions holds:*

- *for all $\lambda\in\Lambda$ all members of s_λ and τ_λ are connected;*
- *for all $\lambda\in\Lambda$ all members of s_λ are connected and all $\lambda\in I$ all t_λ are connected;*
- *for all $\lambda\in\Lambda$ all members of σ_λ and t_λ are connected;*
- *for all $\lambda\in\Lambda$ all members of σ_λ are connected and all $\lambda\in I$ all τ_λ are connected;*

then,

$$\inf_{y\in\mathbb{Y}}\sup_{x\in\mathbb{X}}F(x,y)=\sup_{x\in\mathbb{X}}\inf_{y\in\mathbb{Y}}F(x,y)\ .$$

Lemma 22. *$S_g^{\log}:\mathbb{E}\times\langle\mathbb{B}\rangle\to[0,\infty]$ is lower semicontinuous.*

Proof: It suffices to show that $\{(P,B)\in\mathbb{E}\times\langle\mathbb{B}\rangle|S_g^{\log}(P,B)\leq r\}$ is closed for all $r\in\mathbb{R}$. For $r\in\mathbb{R}$ consider a sequence $(P_t,B_t)_{t\in\mathbb{N}}$ with $\lim_{t\to\infty}(P_t,B_t)=(P,B)$ such that $S_g^{\log}(P_t,B_t)\leq r$ for all t. Then,

$$\begin{aligned}S_g^{\log}(P,B)&=-\sum_{\pi\in\Pi}g(\pi)\sum_{F\in\pi}P(F)\log B(F)\\&=\sum_{\pi\in\Pi}\sum_{\substack{F\in\pi\\ g(\pi)P(F)>0}}-g(\pi)P(F)\log B(F).\end{aligned}$$

If $g(\pi)P(F) > 0$ and $B_t(F)$ converges to zero, then there is an $T \in \mathbb{N}$ such that for all $t \geq T$, $-g(\pi)P_t(F)\log B_t(F) > r+1$. Thus, $B_t(F)$ cannot converge to zero, if $P(F) > 0$. Since (B_t) converges, it has to converge to some $B(F) > 0$. Thus, when $g(\pi)P(F) > 0$ we have that $-g(\pi)P(F)\log B(F) = \lim_{t\to\infty} -g(\pi)P_t(F)\log B_t(F) \leq r$. From $S_g^{\log}(P_t, B_t) \leq r$ we conclude that

$$\begin{aligned}\sum_{\pi\in\Pi}\sum_{\substack{F\in\pi\\ g(\pi)P(F)>0}} -g(\pi)P(F)\log B(F) &= \lim_{t\to\infty}\sum_{\pi\in\Pi}\sum_{\substack{F\in\pi\\ g(\pi)P(F)>0}} -g(\pi)P_t(F)\log B_t(F)\\ &\leq r\end{aligned}$$

□

Proposition 23. *For all* $\mathbb{E}$,

$$\inf_{B\in\langle\mathbb{B}\rangle}\sup_{P\in\mathbb{E}} S_g^{\log}(P,B) = \sup_{P\in\mathbb{E}}\inf_{B\in\langle\mathbb{B}\rangle} S_g^{\log}(P,B) \ .$$

Proof: It suffices to verify that the conditions of Lemma 21 are satisfied.

$\mathbb{E}, \langle\mathbb{B}\rangle$ are subsets of $\mathbb{R}^{|\Omega|}$, $\mathbb{R}^{|\mathscr{P}\Omega|}$ respectively, thus naturally equipped with the induced topology. $\langle\mathbb{B}\rangle$ is compact and Hausdorff (see Lemma 2). $S_g^{\log} : \mathbb{E} \times \langle\mathbb{B}\rangle \to [0,\infty]$ is lower semicontinuous (see Lemma 22).

We need to show that one of the connectivity conditions holds. In fact they all hold, as we shall see.

Note that $\mathbb{E}, \langle\mathbb{B}\rangle$ are connected since they are convex.

For the s_λ and σ_λ consider any $B \in \langle\mathbb{B}\rangle$ and suppose that $P, P' \in \mathbb{E}$ are such that $S_g^{\log}(P,B) \overset{\geq}{>} \lambda$ and $S_g^{\log}(P',B) \overset{\geq}{>} \lambda$. Then for $\eta \in (0,1)$ we have:

$$\begin{aligned}S_g^{\log}(\eta P + (1-\eta)P', B) &= -\sum_{\pi\in\Pi} g(\pi)\sum_{F\in\pi}(\eta P + (1-\eta)P')(F)\log B(F)\\ &= \eta S_g^{\log}(P,B) + (1-\eta)S_g^{\log}(P',B)\\ &\overset{\geq}{>} \lambda\end{aligned} \tag{2}$$

Thus,

$$\{P \in \mathbb{E} \mid S_g^{\log}(P,B) \overset{\geq}{>} \lambda\}$$

is convex for all $B \in \langle\mathbb{B}\rangle$.

Thus, every intersection of such sets is convex. Hence these intersections are connected. (If any such intersection is empty, then it is trivially connected.)

For the t_λ and τ_λ note that for every $P \in \mathbb{P}$ we have that

$$\{B \in \langle\mathbb{B}\rangle \mid S_g^{\log}(P,B) \overset{\leq}{<} \lambda\}$$

is convex, which follows from Proposition 19 by noting that for a convex function (here $S_g^{\log}(P,\cdot)$) on a convex set (here $\langle\mathbb{B}\rangle$), the set of elements in the domain which are mapped to a number (strictly) less than λ is convex for all $\lambda \in \mathbb{R}$.

Thus, every intersection of such sets is convex. Hence these intersections are connected. □

The suprema and infima referred to in Proposition 23 may not be achieved at points of $\mathbb{E}$. If not, they will be achieved instead at points in the closure $[\mathbb{E}]$ of $\mathbb{E}$. We shall use $\arg\sup_{P\in\mathbb{E}}$ (and $\arg\inf_{P\in\mathbb{E}}$) to refer to the points in $[\mathbb{E}]$ that achieve the supremum (respectively infimum) *whether or not these points are in* $\mathbb{E}$.

Theorem 24. *As usual,* $\mathbb{E}$ *is taken to be convex and* g *inclusive. We have that:*

$$\arg\sup_{P\in\mathbb{E}} H_g(P) = \arg\inf_{B\in\mathbb{B}}\sup_{P\in\mathbb{E}} S_g^{\log}(P,B). \tag{3}$$

Proof: We shall prove the following slightly stronger equality allowing B to range in $\langle\mathbb{B}\rangle$ instead of $\mathbb{B}$:

$$\arg\sup_{P\in\mathbb{E}} H_g(P) = \arg\inf_{B\in\langle\mathbb{B}\rangle}\sup_{P\in\mathbb{E}} S_g^{\log}(P,B). \tag{4}$$

The theorem then follows from the following fact. The right hand side of (4) is an optimization problem where the optimum (here we look for the infimum of $\sup_{P\in\mathbb{E}} S_g^{\log}(P,\cdot)$) uniquely obtains for a certain value (here $P^\dagger$). Restricting the domain of the variables (here from $\langle\mathbb{B}\rangle$ to $\mathbb{B}$) in the optimization problem, to a subdomain which contains optimum $P^\dagger \in [\mathbb{E}] \subseteq \mathbb{B} \subseteq \langle\mathbb{B}\rangle$, does not change where the optimum obtains nor the value of the optimum.

Note that,

$$\begin{aligned}\sup_{P\in\mathbb{E}} H_g(P) &= \sup_{P\in\mathbb{E}} S_g^{\log}(P,P)\\ &= \sup_{P\in\mathbb{E}}\inf_{B\in\langle\mathbb{B}\rangle} S_g^{\log}(P,B)\\ &= \inf_{B\in\langle\mathbb{B}\rangle}\sup_{P\in\mathbb{E}} S_g^{\log}(P,B).\end{aligned}$$

The first equality is simply the definition of H_g. The second equality follows directly from strict propriety (Corollary 18). To obtain the third line we apply Proposition 23.

It remains to show that we can introduce arg on both sides of (3).

The following sort of argument seems to be folklore in game theory; we here adapt Grünwald and Dawid (2004, Lemma 4.1, p. 1384) for our purposes. We have

$$\begin{aligned}P^\dagger &:= \arg\sup_{P\in\mathbb{E}} S_g^{\log}(P,P) &(5)\\ &= \arg\sup_{P\in\mathbb{E}}\inf_{B\in\langle\mathbb{B}\rangle} S_g^{\log}(P,B)\ . &(6)\end{aligned}$$

The $\arg\sup$ in (5) is unique (Corollary 10). (6) follows from strict propriety of $S_g^{\log}$ (Corollary 18). Now let

$$B^\dagger \in \arg\inf_{B\in\langle\mathbb{B}\rangle}\sup_{P\in\mathbb{E}} S_g^{\log}(P,B)\ .$$

Then

$$\begin{aligned}S_g^{\log}(P^\dagger,P^\dagger) &= \sup_{P\in\mathbb{E}}\inf_{B\in\langle\mathbb{B}\rangle} S_g^{\log}(P,B) &(7)\\ &= \inf_{B\in\langle\mathbb{B}\rangle} S_g^{\log}(P^\dagger,B)\\ &\le S_g^{\log}(P^\dagger,B^\dagger)\\ &\le \sup_{P\in\mathbb{E}} S_g^{\log}(P,B^\dagger)\\ &= \inf_{B\in\langle\mathbb{B}\rangle}\sup_{P\in\mathbb{E}} S_g^{\log}(P,B). &(8)\end{aligned}$$

The first equality follows from the definition of $P^\dagger$; see (5) and (6). That we may drop the sup again follows from the definition of $P^\dagger$, since $P^\dagger$ maximises $\inf_{B\in\langle\mathbb{B}\rangle} S_g^{\log}(\cdot,B)$. The inequalities hold since dropping a minimisation and introducing a maximisation can only lead to an increase. The final inequality is immediate from the definition of $B^\dagger$ minimising $\sup_{P\in\mathbb{E}} S_g^{\log}(P,\cdot)$.

By Proposition 23 all inequalities above are in fact equalities. From $S_g^{\log}(P^\dagger,P^\dagger) = S_g^{\log}(P^\dagger,B^\dagger)$ and strict propriety we may now infer that $B^\dagger = P^\dagger$. □

In sum, then: if an agent is to minimise her worst-case g-score, then her belief function needs to be the probability function in $\mathbb{E}$ that maximises g-entropy, as long as this entropy maximiser is in $\mathbb{E}$. That the belief function is to be a probability function is the content of the Probability norm; that it is to be in $\mathbb{E}$ is the content of the Calibration norm; that it is to maximise g-entropy is related to the Equivocation norm. We shall defer a full discussion of the Equivocation norm to §4. In the next section we shall show that the arguments of this section generalise to belief as defined over sentences rather than propositions. This will imply that logically equivalent sentences should be believed to the same extent—an important component of the Probability norm in the sentential framework.

We shall conclude this section by providing a slight generalisation of the previous result. Note that thus far when considering worst-case g-score, this worst case is with respect to a chance function taken to be in $\mathbb{E} = \langle\mathbb{P}^*\rangle$. But the evidence determines something more precise, namely that the chance function is in $\mathbb{P}^*$, which is not assumed to be convex. The following result indicates that our main argument will carry over to this more precise setting.

Theorem 25. Suppose $\mathbb{P}^* \subseteq \mathbb{P}$ *is such that the unique* g*-entropy maximiser* $P^\dagger$ *for* $[\mathbb{E}] = [\langle\mathbb{P}^*\rangle]$ *is in* $[\mathbb{P}^*]$*. Then,*

$$P^\dagger = \arg\sup_{P\in\mathbb{E}} H_g(P) = \arg\inf_{B\in\mathbb{B}}\sup_{P\in\mathbb{P}^*} S_g^{\log}(P,B).$$

Proof: As in the previous proof we shall prove a slightly stronger equality:

$$P^\dagger = \arg\sup_{P\in\mathbb{E}} H_g(P) = \arg\inf_{B\in\langle\mathbb{B}\rangle}\sup_{P\in\mathbb{P}^*} S_g^{\log}(P,B).$$

The result follows for the same reasons given in the proof of Theorem 24.

From the strict propriety of $S_g^{\log}$ we have

$$\begin{aligned} S_g^{\log}(P^\dagger,P^\dagger) &= \inf_{B\in\langle\mathbb{B}\rangle} S_g^{\log}(P^\dagger,B) \\ &\le \inf_{B\in\langle\mathbb{B}\rangle}\sup_{P\in\mathbb{P}^*} S_g^{\log}(P,B) \\ &\le \inf_{B\in\langle\mathbb{B}\rangle}\sup_{P\in\langle\mathbb{P}^*\rangle} S_g^{\log}(P,B) \\ &= \sup_{P\in\langle\mathbb{P}^*\rangle} S_g^{\log}(P,P^\dagger) \\ &= S_g^{\log}(P^\dagger,P^\dagger) \end{aligned}$$

where the last two equalities are simply Theorem 24. Hence,

$$\inf_{B\in\langle\mathbb{B}\rangle}\sup_{P\in\mathbb{P}^*} S_g^{\log}(P,B) = S_g^{\log}(P^\dagger,P^\dagger) = \sup_{P\in\mathbb{E}} H_g(P) = \sup_{P\in\mathbb{P}^*} H_g(P).$$

That is, the lowest worst case expected loss is the same for $P \in [\mathbb{P}^*]$ and $P \in [\langle \mathbb{P}^* \rangle]$.

Furthermore, since $S_g^{\log}(P^\dagger, P^\dagger) = \sup_{P \in [\langle \mathbb{P}^* \rangle]} S_g^{\log}(P, P^\dagger)$ and since $P^\dagger \in [\mathbb{P}^*]$ we have $S_g^{\log}(P^\dagger, P^\dagger) = \sup_{P \in \mathbb{P}^*} S_g^{\log}(P, P^\dagger)$. Thus, $B = P^\dagger$ minimises $\sup_{P \in \mathbb{P}^*} S_g^{\log}(P, B)$.

Now suppose that $B' \in \langle \mathbb{B} \rangle$ is different from $P^\dagger$. Then

$$\sup_{P \in \mathbb{P}^*} S_g^{\log}(P, B') \geq S_g^{\log}(P^\dagger, B') > S_g^{\log}(P^\dagger, P^\dagger),$$

where the strict inequality follows from strict propriety. This shows that adopting $B' \neq P^\dagger$ leads to an avoidably bad score.

Hence $B = P^\dagger$ is the unique function in $\langle \mathbb{B} \rangle$ which minimises $\sup_{P \in \mathbb{P}^*} S_g^{\log}(P, B)$. □

§3
Belief over sentences

Armed with our results for beliefs defined over propositions we now tackle the case of beliefs defined over sentences $S\mathcal{L}$ of a propositional language $\mathcal{L}$. The plan is as follows. First we normalise the belief functions in §3.1. In §3.2 we motivate the use of logarithmic loss as a default loss function. We are able to define our logarithmic scoring rule in §3.3, and we show there that, with respect to our scoring rule, the generalised entropy maximiser is the unique belief function that minimises worst-case expected loss.

Again, we shall not impose any restriction—such as additivity—on the agent's belief function, now defined on the sentences of the propositional language $\mathcal{L}$. In particular, we do not assume that the agent's belief function assigns logically equivalent sentences the same degree of belief. We shall show that any belief function violating this property incurs an avoidable loss. Thus the results of this section allow us to show more than we could in the case of belief functions defined over propositions.

Several of the proofs in this section are analogous to the proofs of corresponding results presented in §2. They are included here in full for the sake of completeness; the reader may wish to skim over those details which are already familiar.

§3.1. Normalisation

$S\mathcal{L}$ is the set of sentences of propositional language $\mathcal{L}$, formed as usual by recursively applying the connectives $\neg, \vee, \wedge, \rightarrow, \leftrightarrow$ to the propositional variables $A_1, \ldots, A_n$. A non-normalised belief function $bel : S\mathcal{L} \longrightarrow \mathbb{R}_{\geq 0}$ is thus a function that maps any sentence of the language to a non-negative real number. As in §2.1, for technical convenience we shall focus our attention on normalised belief functions.

Definition 26 (*Representation*). A sentence $\theta \in S\mathcal{L}$ *represents* the proposition $F = \{\omega : \omega \models \theta\}$. Let $\mathscr{F}$ be a set of pairwise distinct propositions. We say that $\Theta \subseteq S\mathcal{L}$ is a set of *representatives* of $\mathscr{F}$, if and only if each sentence in Θ represents some proposition in $\mathscr{F}$ and each proposition in $\mathscr{F}$ is represented by a unique sentence in Θ. A set ρ of representatives of $\mathscr{P}\Omega$ will be called a *representation*. We denote by ϱ the set of all representations. For a set of pairwise distinct propositions $\mathscr{F}$ and a representation $\rho \in \varrho$ we denote by $\rho(\mathscr{F}) \subset S\mathcal{L}$ the set of sentences in ρ which represents the propositions in $\mathscr{F}$.

We call $\pi_{\mathscr{L}} \subseteq S\mathscr{L}$ a *partition of* $S\mathscr{L}$, if and only if it is a set of representatives of some partition $\pi \in \Pi$ of propositions. We denote by $\Pi_{\mathscr{L}}$ the set of these $\pi_{\mathscr{L}}$.

Definition 27 (*Normalised belief function on sentences*). Define the set of normalized belief functions on $S\mathscr{L}$ as

$$\mathbb{B}_{\mathscr{L}} := \{B_{\mathscr{L}} : S\mathscr{L} \longrightarrow [0,1] : \sum_{\varphi\in\pi_{\mathscr{L}}} B_{\mathscr{L}}(\varphi) \leq 1 \text{ for all } \pi_{\mathscr{L}} \in \Pi_{\mathscr{L}} \text{ and } \sum_{\varphi\in\pi_{\mathscr{L}}} B_{\mathscr{L}}(\varphi) = 1 \text{ for some } \pi_{\mathscr{L}} \in \Pi_{\mathscr{L}}\}.$$

The set of probability functions is defined as

$$\mathbb{P}_{\mathscr{L}} := \{P_{\mathscr{L}} : S\mathscr{L} \longrightarrow [0,1] : \sum_{\varphi\in\pi_{\mathscr{L}}} P_{\mathscr{L}}(\varphi) = 1 \text{ for all } \pi_{\mathscr{L}} \in \Pi_{\mathscr{L}}\}.$$

As in the proposition case we have:

Proposition 28. *$P_{\mathscr{L}} \in \mathbb{P}_{\mathscr{L}}$ iff $P_{\mathscr{L}} : S\mathscr{L} \longrightarrow [0,1]$ satisfies the axioms of probability:*

P1: *$P_{\mathscr{L}}(\tau) = 1$ for all tautologies τ.*

P2: *If $\models \neg(\varphi \wedge \psi)$ then $P_{\mathscr{L}}(\varphi \vee \psi) = P_{\mathscr{L}}(\varphi) + P_{\mathscr{L}}(\psi)$.*

Proof: Suppose $P_{\mathscr{L}} \in \mathbb{P}_{\mathscr{L}}$. For any tautology $\tau \in S\mathscr{L}$ it holds that $P_{\mathscr{L}}(\tau) = 1$ because $\{\tau\}$ is a partition in $\Pi_{\mathscr{L}}$. $P_{\mathscr{L}}(\neg\tau) = 0$ because $\{\tau, \neg\tau\}$ is a partition in $\Pi_{\mathscr{L}}$ and $P_{\mathscr{L}}(\tau) = 1$.

Suppose that $\varphi, \psi \in S\mathscr{L}$ are such that $\models \neg(\varphi \wedge \psi)$. We shall proceed by cases to show that $P_{\mathscr{L}}(\varphi \vee \psi) = P_{\mathscr{L}}(\varphi) + P_{\mathscr{L}}(\psi)$. In the first three cases one of the sentences is a contradiction, in the last two cases there are no contradictions.
(i) $\models \varphi$ and $\models \neg\psi$, then $\models \varphi \vee \psi$. Thus by the above $P_{\mathscr{L}}(\varphi) = 1$ and $P_{\mathscr{L}}(\psi) = 0$ and hence $P_{\mathscr{L}}(\varphi \vee \psi) = 1 = P_{\mathscr{L}}(\varphi) + P_{\mathscr{L}}(\psi)$.
(ii) $\models \neg\varphi$ and $\models \neg\psi$, then $\models \neg\varphi \vee \neg\psi$. Thus $P_{\mathscr{L}}(\varphi \vee \psi) = 0 = P_{\mathscr{L}}(\varphi) + P_{\mathscr{L}}(\psi)$.
(iii) $\not\models \neg\varphi$, $\not\models \varphi$, and $\models \neg\psi$, then $\{\varphi \vee \psi, \neg\varphi \vee \psi\}$ and $\{\varphi, \neg\varphi \vee \psi\}$ are both partitions in $\Pi_{\mathscr{L}}$. Thus $P_{\mathscr{L}}(\varphi \vee \psi) + P_{\mathscr{L}}(\neg\varphi \vee \psi) = 1 = P_{\mathscr{L}}(\varphi) + P_{\mathscr{L}}(\neg\varphi \vee \psi)$. Putting these observations together we now find $P_{\mathscr{L}}(\varphi \vee \psi) = P_{\mathscr{L}}(\varphi) = P_{\mathscr{L}}(\varphi) + P_{\mathscr{L}}(\psi)$.
(iv) $\not\models \neg\varphi$, $\not\models \neg\psi$ and $\models \varphi \leftrightarrow \neg\psi$, then $\{\varphi, \psi\}$ is a partition and $\varphi \vee \psi$ is a tautology. Hence, $P_{\mathscr{L}}(\varphi) + P_{\mathscr{L}}(\psi) = 1$ and $P_{\mathscr{L}}(\varphi \vee \psi) = 1$. This now yields $P_{\mathscr{L}}(\varphi) + P_{\mathscr{L}}(\psi) = P_{\mathscr{L}}(\varphi \vee \psi)$.
(v) $\not\models \neg\varphi$, $\not\models \neg\psi$ and $\not\models \varphi \leftrightarrow \neg\psi$, then none of the following sentences is a tautology or a contradiction: $\varphi, \psi, \varphi \vee \psi, \neg(\varphi \vee \psi)$. Since $\{\varphi, \psi, \neg(\varphi \vee \psi)\}$ and $\{\varphi \vee \psi, \neg(\varphi \vee \psi)\}$ are both partitions in $\Pi_{\mathscr{L}}$ we obtain $P_{\mathscr{L}}(\varphi) + P_{\mathscr{L}}(\psi) = 1 - P_{\mathscr{L}}(\neg(\varphi \vee \psi)) = P_{\mathscr{L}}(\varphi \vee \psi)$. So $P_{\mathscr{L}}(\varphi) + P_{\mathscr{L}}(\psi) = P_{\mathscr{L}}(\varphi \vee \psi)$.

On the other hand, suppose P1 and P2 hold. That $\sum_{\varphi\in\pi_{\mathscr{L}}} P_{\mathscr{L}}(\varphi) = 1$ holds for all $\pi_{\mathscr{L}} \in \Pi_{\mathscr{L}}$ can be seen by induction on the size of $\pi_{\mathscr{L}}$. If $|\pi_{\mathscr{L}}| = 1$ then $\pi = \{\tau\}$ for some tautology $\tau \in S\mathscr{L}$ and $P_{\mathscr{L}}(\tau) = 1$ by P1. Suppose then that $\pi_{\mathscr{L}} = \{\varphi_1, \ldots, \varphi_{k+1}\}$ for $k \geq 1$. Now $\sum_{i=1}^{k-1} P_{\mathscr{L}}(\varphi_i) + P_{\mathscr{L}}(\varphi_k \vee \varphi_{k+1}) = 1$ by the induction hypothesis. Furthermore, $P_{\mathscr{L}}(\varphi_k \vee \varphi_{k+1}) = P_{\mathscr{L}}(\varphi_k) + P_{\mathscr{L}}(\varphi_{k+1})$ by P2, so $\sum_{\varphi\in\pi_{\mathscr{L}}} P_{\mathscr{L}}(\varphi) = 1$ as required. □

Definition 29 (*Respects logical equivalence*). We say that a belief function $B_{\mathscr{L}} \in \langle\mathbb{B}_{\mathscr{L}}\rangle$ *respects logical equivalence* if and only if $\models \varphi \leftrightarrow \psi$ implies $B_{\mathscr{L}}(\varphi) = B_{\mathscr{L}}(\psi)$.

Proposition 30. *The probability functions $P_{\mathscr{L}} \in \mathbb{P}_{\mathscr{L}}$ respect logical equivalence.*

Proof: Suppose $P_{\mathscr{L}} \in \mathbb{P}_{\mathscr{L}}$ and assume that $\varphi, \psi \in S\mathscr{L}$ are logically equivalent. Note that $\psi \wedge \neg\varphi \vDash A_1 \wedge \neg A_1$, $\psi \vee \neg\varphi \vDash A_1 \vee \neg A_1$ and that $\{\varphi, \neg\varphi\}$ and $\{\psi, \neg\varphi\}$ are partitions in $\Pi_{\mathscr{L}}$. Hence,

$$P_{\mathscr{L}}(\varphi) + P_{\mathscr{L}}(\neg\varphi) = 1 = P_{\mathscr{L}}(\psi) + P_{\mathscr{L}}(\neg\varphi).$$

Therefore, $P_{\mathscr{L}}(\varphi) = P_{\mathscr{L}}(\psi)$.

Thus, the $P_{\mathscr{L}} \in \mathbb{P}_{\mathscr{L}}$ assign logically equivalent formulae the same probability. □

§3.2. Loss

By analogy with the line of argument of §2.3, we shall suppose that a default loss function $L : S\mathscr{L} \times \langle \mathbb{B}_{\mathscr{L}} \rangle \to (-\infty, \infty]$ satisfies the following requirements:

L1. $L(\varphi, B_{\mathscr{L}}) = 0$, if $B_{\mathscr{L}}(\varphi) = 1$.

L2. $L(\varphi, B_{\mathscr{L}})$ strictly increases as $B_{\mathscr{L}}(\varphi)$ decreases from 1 towards 0.

L3. $L(\varphi, B_{\mathscr{L}})$ only depends on $B_{\mathscr{L}}(\varphi)$.

Suppose we have a fixed belief function $B_{\mathscr{L}} \in \langle \mathbb{B}_{\mathscr{L}} \rangle$ such that $B_{\mathscr{L}}(\tau) = 1$ for any tautology τ, and $\mathscr{L} = \mathscr{L}_1 \cup \mathscr{L}_2$ where $\mathscr{L}_1$ and $\mathscr{L}_2$ are ***independent sublanguages***, written $\mathscr{L}_1 \perp\!\!\!\perp_{B_{\mathscr{L}}} \mathscr{L}_2$, i.e., $B_{\mathscr{L}}(\phi_1 \wedge \phi_2) = B_{\mathscr{L}}(\phi_1) \cdot B_{\mathscr{L}}(\phi_2)$ for all $\phi_1 \in S\mathscr{L}_1$ and $\phi_2 \in S\mathscr{L}_2$. Let $B_{\downharpoonright \mathscr{L}_1}(\phi_1) := B_{\mathscr{L}}(\phi_1)$, $B_{\downharpoonright \mathscr{L}_2}(\phi_2) := B_{\mathscr{L}}(\phi_2)$.

L4. Losses are additive when the language is composed of independent sublanguages: if $\mathscr{L} = \mathscr{L}_1 \cup \mathscr{L}_2$ for $\mathscr{L}_1 \perp\!\!\!\perp_{B_{\mathscr{L}}} \mathscr{L}_2$ then $L(\phi_1 \wedge \phi_2, B_{\mathscr{L}}) = L_1(\phi_1, B_{\downharpoonright \mathscr{L}_1}) + L_2(\phi_2, B_{\downharpoonright \mathscr{L}_2})$, where L_1, L_2 are loss functions defined on $\mathscr{L}_1, \mathscr{L}_2$ respectively.

Theorem 31. *If a loss function L on $S\mathscr{L} \times \langle \mathbb{B}_{\mathscr{L}} \rangle$ satisfies L1–4, then $L(\varphi, B_{\mathscr{L}}) = -k \log B_{\mathscr{L}}(\varphi)$, where the constant $k > 0$ does not depend on the language $\mathscr{L}$.*

Proof: We shall first focus on a loss function L defined with respect to a language $\mathscr{L}$ that contains at least two propositional variables.

L3 implies that $L(\varphi, B_{\mathscr{L}}) = f_{\mathscr{L}}(B_{\mathscr{L}}(\varphi))$ for some function $f_{\mathscr{L}} : [0,1] \longrightarrow (-\infty, \infty]$. For our fixed $\mathscr{L}$ and all $x, y \in [0,1]$ choose some $B_{\mathscr{L}} \in \langle \mathbb{B}_{\mathscr{L}} \rangle$ such that $\mathscr{L} = \mathscr{L}_1 \cup \mathscr{L}_2$, $\mathscr{L}_1 \perp\!\!\!\perp_{B_{\mathscr{L}}}, \mathscr{L}_2 B_{\mathscr{L}}(\phi_1) = x$, and $B_{\mathscr{L}}(\phi_2) = y$ for some $\phi_1 \in S\mathscr{L}_1$, $\phi_2 \in S\mathscr{L}_2$. This is possible because $\mathscr{L}$ contains at least two propositional variables.

Note that since $\mathscr{L}_1$ and $\mathscr{L}_2$ are independent sublanguages, given some specific tautology τ_1 of $\mathscr{L}_1$,

$$1 = B_{\mathscr{L}}(\tau_1) = B_{\downharpoonright \mathscr{L}_1}(\tau_1). \tag{9}$$

$B_{\mathscr{L}}(\tau_1)$ is well defined since τ_1 is a tautology of $S\mathscr{L}_1$ and every sentence in $S\mathscr{L}_1$ is a sentence in $S\mathscr{L}$. Similarly, $B_{\downharpoonright \mathscr{L}_2}(\tau_2) = 1$ for some specific tautology τ_2 of $\mathscr{L}_2$. By L1, then, $L_1(\tau_1, B_{\downharpoonright \mathscr{L}_1}) = L_2(\tau_2, B_{\downharpoonright \mathscr{L}_2}) = 0$, where L_1, respectively L_2, are loss functions with respect to $S\mathscr{L}_1$ and $S\mathscr{L}_2$ satisfying L1–4. Thus,

$$\begin{array}{rcl}
f_{\mathscr{L}}(x\cdot y) & = & f_{\mathscr{L}}(B_{\mathscr{L}}(\phi_1)\cdot B_{\mathscr{L}}(\phi_2)) \\
& \overset{L_3}{=} & L(\phi_1\wedge\phi_2, B_{\mathscr{L}}) \\
& \overset{L4}{=} & L_1(\phi_1, B_{\downharpoonleft\mathscr{L}_1}) + L_2(\phi_2, B_{\downharpoonleft\mathscr{L}_2}) \\
& \overset{L4}{=} & [L(\phi_1\wedge\tau_2, B_{\mathscr{L}}) - L_2(\tau_2, B_{\downharpoonleft\mathscr{L}_2})] \\
& & \quad + [L(\tau_1\wedge\phi_2, B_{\mathscr{L}}) - L_1(\tau_1, B_{\downharpoonleft\mathscr{L}_1})] \\
& \overset{L1}{=} & L(\phi_1\wedge\tau_2, B_{\mathscr{L}}) + L(\tau_1\wedge\phi_2, B_{\mathscr{L}}) \\
& \overset{L3}{=} & f_{\mathscr{L}}(B_{\mathscr{L}}(\phi_1\wedge\tau_2)) + f_{\mathscr{L}}(B_{\mathscr{L}}(\phi_2\wedge\tau_1)) \\
& = & f_{\mathscr{L}}(B_{\downharpoonleft\mathscr{L}_1}(\phi_1)\cdot B_{\downharpoonleft\mathscr{L}_2}(\tau_2)) + f_{\mathscr{L}}(B_{\downharpoonleft\mathscr{L}_1}(\tau_1)\cdot B_{\downharpoonleft\mathscr{L}_2}(\phi_2)) \\
& \overset{(9)}{=} & f_{\mathscr{L}}(B_{\downharpoonleft\mathscr{L}_1}(\phi_1)) + f_{\mathscr{L}}(B_{\downharpoonleft\mathscr{L}_2}(\phi_2)) \\
& = & f_{\mathscr{L}}(B_{\mathscr{L}}(\phi_1)) + f_{\mathscr{L}}(B_{\mathscr{L}}(\phi_2)) \\
& = & f_{\mathscr{L}}(x) + f_{\mathscr{L}}(y).
\end{array}$$

The negative logarithm on $(0,1]$ is characterisable up to a multiplicative constant $k_{\mathscr{L}}$ in terms of this additivity, together with the condition that $f_{\mathscr{L}}(x)\geq 0$ which is implied by L1–2 (see, e.g., Aczél and Daróczy, 1975, Theorem 0.2.5). L2 ensures that $f_{\mathscr{L}}$ is not zero everywhere, so $k_{\mathscr{L}} > 0$. As in the corresponding proof for propositions, it follows that $f_{\mathscr{L}}(0) = +\infty$.

Thus far we have shown that for a fixed language $\mathscr{L}$ with at least two propositional variables, $L(F, B_{\mathscr{L}}) = -k_{\mathscr{L}}\log B_{\mathscr{L}}(F)$ on $[0,1]$.

Now focus on an arbitrary language $\mathscr{L}_1$ and a corresponding loss function L_1. We can choose $\mathscr{L}_2, \mathscr{L}, B_{\mathscr{L}}$ such that $\mathscr{L}$ is composed of independent sublanguages $\mathscr{L}_1$ and $\mathscr{L}_2$. By reasoning analogous to that above,

$$\begin{array}{rcl}
f_{\mathscr{L}_1}(B_{\downharpoonleft\mathscr{L}_1}(\phi_1)) & = & L_1(\phi_1, B_{\downharpoonleft\mathscr{L}_1}) \\
& = & L(\phi_1\wedge\tau_2, B_{\mathscr{L}}) \\
& = & f_{\mathscr{L}}(B_{\mathscr{L}}(\phi_1\wedge\tau_2)) \\
& = & f_{\mathscr{L}}(B_{\mathscr{L}}(\phi_1)\cdot 1) \\
& = & -k_{\mathscr{L}}\log B_{\downharpoonleft\mathscr{L}_1}(\phi_1).
\end{array}$$

So the loss function for $\mathscr{L}_1$ is $L_1(\phi_1, B_{\downharpoonleft\mathscr{L}_1}) = -k_{\mathscr{L}}\log B_{\downharpoonleft\mathscr{L}_1}(\phi_1)$. Thus the constant $k_{\mathscr{L}}$ does not depend on $\mathscr{L}$ after all.

In general, then, $L(F, B_{\mathscr{L}}) = -k\log B_{\mathscr{L}}(F)$ for some positive k. □

Since multiplication by a constant is equivalent to change of base, we can take log to be the natural logarithm. Since we will be interested in the belief functions that minimise loss, rather than in the absolute value of any particular losses, we can take $k = 1$ without loss of generality. Theorem 31 thus allows us to focus on the logarithmic loss function:

$$L^{\log}(F, B_{\mathscr{L}}) := -\log B_{\mathscr{L}}(F).$$

§3.3. Score, entropy and their connection

In the case of belief over sentences, the expected loss varies according to which sentences are used to represent the various partitions of propositions. We can define the g-score to be the worst-case expected loss, where this worst case is taken over all possible representations:

Definition 32 (*g-score*). Given a loss function L, an inclusive weighting function $g : \Pi \longrightarrow \mathbb{R}_{\geq 0}$ and a representation $\rho \in \varrho$ we define the *representation-relative g-score* $S^L_{g,\rho} : \mathbb{P}_{\mathscr{L}} \times \langle \mathbb{B}_{\mathscr{L}} \rangle \longrightarrow [-\infty, \infty]$ by

$$S^L_{g,\rho}(P_{\mathscr{L}}, B_{\mathscr{L}}) := \sum_{\pi \in \Pi} g(\pi) \sum_{\varphi \in \rho(\pi)} P_{\mathscr{L}}(\varphi) L(\varphi, B_{\mathscr{L}}),$$

and the *(representation-independent) g-score* $S^L_{g,\mathscr{L}} : \mathbb{P}_{\mathscr{L}} \times \langle \mathbb{B}_{\mathscr{L}} \rangle \longrightarrow [-\infty, \infty]$ by

$$S^L_{g,\mathscr{L}}(P_{\mathscr{L}}, B_{\mathscr{L}}) := \sup_{\rho \in \varrho} S^L_{g,\rho}(P_{\mathscr{L}}, B_{\mathscr{L}}).$$

In particular, for the logarithmic loss function under consideration here, we have,

$$S^{\log}_{g,\rho}(P_{\mathscr{L}}, B_{\mathscr{L}}) := -\sum_{\pi \in \Pi} g(\pi) \sum_{\varphi \in \rho(\pi)} P_{\mathscr{L}}(\varphi) \log B_{\mathscr{L}}(\varphi),$$

and

$$S^{\log}_{g,\mathscr{L}}(P_{\mathscr{L}}, B_{\mathscr{L}}) := \sup_{\rho \in \varrho} S^{\log}_{g,\rho}(P_{\mathscr{L}}, B_{\mathscr{L}}).$$

We can thus define the g-entropy of a belief function on $S\mathscr{L}$ as

$$H_{g,\mathscr{L}}(B_{\mathscr{L}}) := S^{\log}_{g,\mathscr{L}}(B_{\mathscr{L}}, B_{\mathscr{L}}).$$

There is a canonical one-to-one correspondence between the $B_{\mathscr{L}} \in \langle \mathbb{B}_{\mathscr{L}} \rangle$ which respect logical equivalence and the $B \in \langle \mathbb{B} \rangle$. In particular, $\mathbb{P}_{\mathscr{L}}$ can be identified with $\mathbb{P}$. Moreover, any convex $\mathbb{E} \subseteq \mathbb{P}$ is in one-to-one correspondence with a convex $\mathbb{E}_{\mathscr{L}} \subseteq \mathbb{P}_{\mathscr{L}}$. In the following we shall make frequent use of this correspondence. For a $B_{\mathscr{L}} \in \langle \mathbb{B}_{\mathscr{L}} \rangle$ which respects logical equivalence we denote by B the function in $\langle \mathbb{B} \rangle$ with which it stands in one-to-one correspondence.

Lemma 33. *If $B_{\mathscr{L}} \in \langle \mathbb{B}_{\mathscr{L}} \rangle$ respects logical equivalence, then for all $\rho \in \varrho$ we have $S^{\log}_{g,\mathscr{L}}(P_{\mathscr{L}}, B_{\mathscr{L}}) = \sup_{\rho \in \varrho} S^{\log}_{g,\rho}(P_{\mathscr{L}}, B_{\mathscr{L}}) = S^{\log}_g(P, B)$.*

Proof: Simply note that $S^{\log}_{g,\rho}(P_{\mathscr{L}}, B_{\mathscr{L}})$ does not depend on ρ. □

Lemma 34. *For all convex $\mathbb{E}_{\mathscr{L}} \subseteq \mathbb{P}_{\mathscr{L}}$,*

$$B^{\dagger}_{\mathscr{L}} \in \arg \inf_{B_{\mathscr{L}} \in \langle \mathbb{B}_{\mathscr{L}} \rangle} \sup_{P_{\mathscr{L}} \in \mathbb{E}_{\mathscr{L}}} \sup_{\rho \in \varrho} S^{\log}_{g,\rho}(P_{\mathscr{L}}, B_{\mathscr{L}})$$

respects logical equivalence.

Proof: Suppose that

$$B^{\dagger}_{\mathscr{L}} \in \arg \inf_{B_{\mathscr{L}} \in \langle \mathbb{B}_{\mathscr{L}} \rangle} \sup_{P_{\mathscr{L}} \in \mathbb{E}_{\mathscr{L}}} \sup_{\rho \in \varrho} S^{\log}_{g,\rho}(P_{\mathscr{L}}, B_{\mathscr{L}}) \tag{10}$$

and assume that $B^{\dagger}_{\mathscr{L}}$ does not respect logical equivalence. Then define

$$B^{\inf}_{\mathscr{L}}(\varphi) := \inf_{\substack{\theta \in S\mathscr{L} \\ \models \theta \leftrightarrow \varphi}} B^{\dagger}_{\mathscr{L}}(\theta)\ . \tag{11}$$

Since $B^{\dagger}_{\mathscr{L}}$ does not respect logical equivalence, there are logically equivalent φ,ψ such that $B^{\dagger}_{\mathscr{L}}(\varphi) \neq B^{\dagger}_{\mathscr{L}}(\psi)$. Hence, $B^{\inf}_{\mathscr{L}}(\varphi) < \max\{B^{\dagger}_{\mathscr{L}}(\varphi), B^{\dagger}_{\mathscr{L}}(\psi)\}$. Thus, for every $\pi_{\mathscr{L}} \in \Pi_{\mathscr{L}}$ with $\varphi \in \pi_{\mathscr{L}}$ we have $\sum_{\chi\in\pi_{\mathscr{L}}} B^{\inf}_{\mathscr{L}}(\chi) < 1$. Thus, $B^{\inf}_{\mathscr{L}} \notin \mathbb{P}_{\mathscr{L}}$. $B^{\inf}_{\mathscr{L}}$ respects logical equivalence by definition.

Now consider the function $B^{\inf} : \mathscr{P}\Omega \longrightarrow [0,1]$ which is determined by $B^{\inf}_{\mathscr{L}}$. Clearly, $B^{\inf} \notin \mathbb{P}$. There are two cases to consider.

(a) $B^{\inf} \in \langle\mathbb{B}\rangle \setminus \mathbb{P}$. Since $B^{\inf} \notin \mathbb{P}$, by Theorem 24 we have that

$$\sup_{P\in\mathbb{E}} S^{\log}_{g}(P,B^{\inf}) > \inf_{B\in\langle\mathbb{B}\rangle} \sup_{P\in\mathbb{E}} S^{\log}_{g}(P,B). \tag{12}$$

(b) $B^{\inf} \notin \langle\mathbb{B}\rangle$. Then define B' by $B'(F) := B^{\inf}(F) + \delta$ for all $F \subseteq \Omega$, where $\delta \in (0,1]$ is minimal such that $B' \in \langle\mathbb{B}\rangle$. In particular $B'(\emptyset) \geq \delta > 0$, thus $B' \notin \mathbb{P}$. Moreover, whenever $P(F) > 0$ it holds that $-P(F)\log B^{\inf}(F) > -P(F)\log B'(F) < +\infty$. For the remainder of this proof we shall extend the definition of the logarithmic g-score $S^{\log}_{g}(P,B)$ by allowing the belief function B to be any non-negative function defined on $\mathscr{P}\Omega$, rather than just $B \in \langle\mathbb{B}\rangle$—if $B \notin \langle\mathbb{B}\rangle$ we shall be careful not to appeal to results that assume $B \in \langle\mathbb{B}\rangle$. We thus find for all $P \in \mathbb{P}$ that $S^{\log}_{g}(P,B^{\inf}) > S^{\log}_{g}(P,B') < +\infty$. Thus, by Theorem 24 we obtain the sharp inequality in the following

$$\begin{aligned}\sup_{P\in\mathbb{E}} S^{\log}_{g}(P,B^{\inf}) &\geq \sup_{P\in\mathbb{E}} S^{\log}_{g}(P,B') \\ &> \inf_{B\in\langle\mathbb{B}\rangle} \sup_{P\in\mathbb{E}} S^{\log}_{g}(P,B). \qquad (13)\end{aligned}$$

For both cases we will obtain a contradiction:

$$\begin{aligned}
S^{\log}_{g}(P^{\dagger},P^{\dagger}) &= \sup_{P\in\mathbb{E}} S^{\log}_{g}(P,P^{\dagger}) && (14)\\
&= \sup_{P_{\mathscr{L}}\in\mathbb{E}_{\mathscr{L}}} \sup_{\rho\in\varrho} S^{\log}_{g,\rho}(P_{\mathscr{L}},P^{\dagger}_{\mathscr{L}}) && (15)\\
&\geq \inf_{B_{\mathscr{L}}\in\langle\mathbb{B}_{\mathscr{L}}\rangle} \sup_{P_{\mathscr{L}}\in\mathbb{E}_{\mathscr{L}}} \sup_{\rho\in\varrho} S^{\log}_{g,\rho}(P_{\mathscr{L}},B_{\mathscr{L}}) && (16)\\
&\overset{(10)}{=} \sup_{P_{\mathscr{L}}\in\mathbb{E}_{\mathscr{L}}} \sup_{\rho\in\varrho} S^{\log}_{g,\rho}(P_{\mathscr{L}},B^{\dagger}_{\mathscr{L}}) && (17)\\
&= \sup_{P_{\mathscr{L}}\in\mathbb{E}_{\mathscr{L}}} -\sum_{\pi\in\Pi} g(\pi) \sum_{\varphi\in\rho(\pi)} P_{\mathscr{L}}(\varphi) \inf_{\substack{\theta\in S\mathscr{L}\\ \vDash\theta\leftrightarrow\varphi}} \log B^{\dagger}_{\mathscr{L}}(\theta) \quad \text{for all } \rho\in\varrho && (18)\\
&= \sup_{P_{\mathscr{L}}\in\mathbb{E}_{\mathscr{L}}} S^{\log}_{g,\rho}(P_{\mathscr{L}},B^{\inf}_{\mathscr{L}}) \quad \text{for all } \rho\in\varrho && (19)\\
&= \sup_{P\in\mathbb{E}} S^{\log}_{g}(P,B^{\inf}) && (20)\\
&> \inf_{B\in\langle\mathbb{B}\rangle} \sup_{P\in\mathbb{E}} S^{\log}_{g}(P,B) && (21)\\
&= S^{\log}_{g}(P^{\dagger},P^{\dagger}). && (22)
\end{aligned}$$

We obtain (14) by noticing that $P^{\dagger}$ is the unique function minimising worst-case g-expected loss (Theorem 24) and recalling that (7)=(8).

(15) is immediate as the probability functions respect logical equivalence. For (18) note that $P_{\mathscr{L}}$ respects logical equivalence. Furthermore, since $-\log(\cdot)$ is strictly decreasing, a smaller value of $B_{\mathscr{L}}(\varphi)$ leads to a greater score.

(19) follows from (11) and Lemma 33 since $B^{\inf}_{\mathscr{L}}$ respects logical equivalence. Hence $S^{\log}_{g,\rho}(P, B^{\inf}_{\mathscr{L}})$ does not depend on the partition ρ.

The inequality (21) we have seen above in the two cases (12) and (13). (22) is again implied by Theorem 24.

We have thus found a contradiction. Hence, the

$$B^{\dagger}_{\mathscr{L}} \in \arg \inf_{B_{\mathscr{L}} \in \langle \mathbb{B}_{\mathscr{L}} \rangle} \sup_{P_{\mathscr{L}} \in \mathbb{E}_{\mathscr{L}}} \sup_{\rho \in \varrho} S^{\log}_{g,\rho}(P_{\mathscr{L}}, B_{\mathscr{L}})$$

have to respect logical equivalence. □

Theorem 24, the key result in the case of belief over propositions, generalises to the case of belief over sentences:

Theorem 35. *As usual, $\mathbb{E}_{\mathscr{L}} \subseteq \mathbb{P}_{\mathscr{L}}$ is taken to be convex and g inclusive. We have that:*

$$\arg \sup_{P_{\mathscr{L}} \in \mathbb{E}_{\mathscr{L}}} H_{g,\mathscr{L}}(P_{\mathscr{L}}) = \arg \inf_{B_{\mathscr{L}} \in \mathbb{B}_{\mathscr{L}}} \sup_{P_{\mathscr{L}} \in \mathbb{E}_{\mathscr{L}}} S^{\log}_{g,\mathscr{L}}(P_{\mathscr{L}}, B_{\mathscr{L}}).$$

Proof: As in the corresponding theorem for proposition (Theorem 24) we shall prove a slightly stronger equality:

$$\arg \sup_{P_{\mathscr{L}} \in \mathbb{E}_{\mathscr{L}}} H_{g,\mathscr{L}}(P_{\mathscr{L}}) = \arg \inf_{B_{\mathscr{L}} \in \langle \mathbb{B}_{\mathscr{L}} \rangle} \sup_{P_{\mathscr{L}} \in \mathbb{E}_{\mathscr{L}}} S^{\log}_{g,\mathscr{L}}(P_{\mathscr{L}}, B_{\mathscr{L}}).$$

Theorem 35 then follows for the same reasons given in the previous section.

Denote by $\langle \mathbb{B}^{le}_{\mathscr{L}} \rangle \subset \langle \mathbb{B}_{\mathscr{L}} \rangle$ the convex hull of functions $B_{\mathscr{L}} \in \mathbb{B}_{\mathscr{L}}$ which respect logical equivalence. Let $Rep : \langle \mathbb{B} \rangle \longrightarrow \langle \mathbb{B}^{le}_{\mathscr{L}} \rangle$ be the bijective map which assigns to any $B \in \langle \mathbb{B} \rangle$ the unique $B_{\mathscr{L}} \in \langle \mathbb{B}_{\mathscr{L}} \rangle$ which represents it (i.e., $B(F) = B_{\mathscr{L}}(\varphi)$, whenever $F \subseteq \Omega$ is represented by $\varphi \in S\mathscr{L}$).

$$\arg \inf_{B_{\mathscr{L}} \in \langle \mathbb{B}_{\mathscr{L}} \rangle} \sup_{P_{\mathscr{L}} \in \mathbb{E}_{\mathscr{L}}} S^{\log}_{g,\mathscr{L}}(P_{\mathscr{L}}, B_{\mathscr{L}}) = \arg \inf_{B_{\mathscr{L}} \in \langle \mathbb{B}^{le}_{\mathscr{L}} \rangle} \sup_{P_{\mathscr{L}} \in \mathbb{E}_{\mathscr{L}}} S^{\log}_{g,\mathscr{L}}(P_{\mathscr{L}}, B_{\mathscr{L}}) \tag{23}$$

$$= Rep(\arg \inf_{B \in \mathbb{B}} \sup_{P \in \mathbb{E}} S^{\log}_{g}(P, B)) \tag{24}$$

$$= Rep(P^{\dagger}) \tag{25}$$

$$= P^{\dagger}_{\mathscr{L}}. \tag{26}$$

(23) is simply Lemma 34. (24) follows directly from applying Lemma 33 and (25) is simply Theorem 24. □

In the above we used $P^{\dagger}_{\mathscr{L}}$ to denote the probability function in $\mathbb{E}_{\mathscr{L}}$ which represents the g-entropy maximiser $P^{\dagger} \in \mathbb{E}$. Now note that $H_{g,\mathscr{L}}(P_{\mathscr{L}}) = H_g(P)$. Thus $P^{\dagger}_{\mathscr{L}}$ is not only the function representing $P^{\dagger}$, it is also the unique function in $\mathbb{E}_{\mathscr{L}}$ which maximises g-entropy $H_{g,\mathscr{L}}$.

Theorem 25 also extends to the sentence framework. As we shall now see, the worst case g-score can be taken with respect to a chance function in $\mathbb{P}^{*}_{\mathscr{L}}$, rather than $\mathbb{E}_{\mathscr{L}} = \langle \mathbb{P}^{*}_{\mathscr{L}} \rangle$.

Theorem 36. *If $\mathbb{P}^{*}_{\mathscr{L}} \subseteq \mathbb{P}_{\mathscr{L}}$ is such that the unique g-entropy maximiser $P^{\dagger}_{\mathscr{L}}$ of $[\mathbb{E}_{\mathscr{L}}] = [\langle \mathbb{P}^{*}_{\mathscr{L}} \rangle]$ is in $[\mathbb{P}^{*}_{\mathscr{L}}]$, then*

$$P^{\dagger}_{\mathscr{L}} = \arg \sup_{P_{\mathscr{L}} \in \mathbb{E}_{\mathscr{L}}} H_{g,\mathscr{L}}(P_{\mathscr{L}}) = \arg \inf_{B \in \mathbb{B}_{\mathscr{L}}} \sup_{P_{\mathscr{L}} \in \mathbb{P}^{*}_{\mathscr{L}}} S^{\log}_{g,\mathscr{L}}(P_{\mathscr{L}}, B_{\mathscr{L}}).$$

Proof: Again, we shall prove a slightly stronger statement with $B_{\mathscr{L}}$ ranging in $\langle \mathbb{B}_{\mathscr{L}} \rangle$.

Since g is inclusive, we have that $S^{\log}_{g}$ is a strictly proper scoring rule. Hence, for a fixed $\rho \in \varrho$, $S^{\log}_{g,\rho}(P_{\mathscr{L}},\cdot)$ is minimal if and only if $P_{\mathscr{L}}(\varphi) = B_{\mathscr{L}}(\varphi)$ for all $\varphi \in \rho$.

Now suppose $B_{\mathscr{L}} \in \langle \mathbb{B}_{\mathscr{L}} \rangle$ is different from a fixed $P_{\mathscr{L}} \in \mathbb{P}_{\mathscr{L}}$. Then there is some $\varphi \in S\mathscr{L}$ such that $B_{\mathscr{L}}(\varphi) \neq P_{\mathscr{L}}(\varphi)$. Now pick some $\rho' \in \varrho$ such that $\varphi \in \rho'$. Then strict propriety implies the sharp inequality below

$$\begin{aligned}
S^{\log}_{g,\mathscr{L}}(P_{\mathscr{L}},B_{\mathscr{L}}) &= \sup_{\rho\in\varrho} S^{\log}_{g,\rho}(P_{\mathscr{L}},B_{\mathscr{L}}) \\
&\geq S^{\log}_{g,\rho'}(P_{\mathscr{L}},B_{\mathscr{L}}) \\
&> S^{\log}_{g,\rho'}(P_{\mathscr{L}},P_{\mathscr{L}}) \\
&= \sup_{\rho\in\varrho} S^{\log}_{g,\rho}(P_{\mathscr{L}},P_{\mathscr{L}}) \\
&= S^{\log}_{g,\mathscr{L}}(P_{\mathscr{L}},P_{\mathscr{L}}).
\end{aligned}$$

The second equality follows since the $P_{\mathscr{L}} \in \mathbb{P}_{\mathscr{L}}$ respect logical equivalence and hence $S^{L}_{g,\rho}(P_{\mathscr{L}},P_{\mathscr{L}})$ does not depend on ρ. Thus, for all $P_{\mathscr{L}} \in \mathbb{P}_{\mathscr{L}}$ we find $\arg\inf_{B_{\mathscr{L}}\in\langle\mathbb{B}_{\mathscr{L}}\rangle} S^{\log}_{g}(P_{\mathscr{L}},B_{\mathscr{L}}) = P_{\mathscr{L}}$. Hence for $P_{\mathscr{L}} = P^{\dagger}_{\mathscr{L}}$ we obtain

$$\begin{aligned}
S^{\log}_{g,\mathscr{L}}(P^{\dagger}_{\mathscr{L}},P^{\dagger}_{\mathscr{L}}) &= \inf_{B_{\mathscr{L}}\in\langle\mathbb{B}_{\mathscr{L}}\rangle} S^{\log}_{g,\mathscr{L}}(P^{\dagger}_{\mathscr{L}},B_{\mathscr{L}}) \\
&\leq \inf_{B_{\mathscr{L}}\in\langle\mathbb{B}_{\mathscr{L}}\rangle} \sup_{P_{\mathscr{L}}\in\mathbb{P}^{*}_{\mathscr{L}}} S^{\log}_{g,\mathscr{L}}(P_{\mathscr{L}},B_{\mathscr{L}}) \\
&\leq \inf_{B_{\mathscr{L}}\in\langle\mathbb{B}_{\mathscr{L}}\rangle} \sup_{P_{\mathscr{L}}\in\langle\mathbb{P}^{*}_{\mathscr{L}}\rangle} S^{\log}_{g,\mathscr{L}}(P_{\mathscr{L}},B_{\mathscr{L}}) \\
&= \sup_{P_{\mathscr{L}}\in\langle\mathbb{P}^{*}_{\mathscr{L}}\rangle} S^{\log}_{g,\mathscr{L}}(P_{\mathscr{L}},P^{\dagger}_{\mathscr{L}}) \\
&= S^{\log}_{g,\mathscr{L}}(P^{\dagger}_{\mathscr{L}},P^{\dagger}_{\mathscr{L}})
\end{aligned}$$

where the last two equalities are simply Theorem 35. Hence,

$$\inf_{B_{\mathscr{L}}\in\langle\mathbb{B}_{\mathscr{L}}\rangle} \sup_{P_{\mathscr{L}}\in\mathbb{P}^{*}_{\mathscr{L}}} S^{\log}_{g,\mathscr{L}}(P_{\mathscr{L}},B_{\mathscr{L}}) = S^{\log}_{g,\mathscr{L}}(P^{\dagger}_{\mathscr{L}},P^{\dagger}_{\mathscr{L}}) = \sup_{P_{\mathscr{L}}\in\langle\mathbb{P}^{*}_{\mathscr{L}}\rangle} H_{g,\mathscr{L}}(P).$$

That is, the lowest worst-case expected loss is the same for $P_{\mathscr{L}} \in [\mathbb{P}^{*}_{\mathscr{L}}]$ and $P_{\mathscr{L}} \in [\langle\mathbb{P}^{*}_{\mathscr{L}}\rangle]$.

Furthermore, since $S^{\log}_{g,\mathscr{L}}(P^{\dagger}_{\mathscr{L}},P^{\dagger}_{\mathscr{L}}) = \sup_{P_{\mathscr{L}}\in\langle\mathbb{P}^{*}_{\mathscr{L}}\rangle} S^{\log}_{g,\mathscr{L}}(P_{\mathscr{L}},P^{\dagger}_{\mathscr{L}})$ and since $P^{\dagger}_{\mathscr{L}} \in [\mathbb{P}^{*}_{\mathscr{L}}]$ we have $S^{\log}_{g,\mathscr{L}}(P^{\dagger}_{\mathscr{L}},P^{\dagger}_{\mathscr{L}}) = \sup_{P_{\mathscr{L}}\in\mathbb{P}^{*}_{\mathscr{L}}} S^{\log}_{g,\mathscr{L}}(P_{\mathscr{L}},P^{\dagger}_{\mathscr{L}})$. Thus, $B_{\mathscr{L}} = P^{\dagger}_{\mathscr{L}}$ minimises $\sup_{P_{\mathscr{L}}\in\mathbb{P}^{*}_{\mathscr{L}}} S^{\log}_{g,\mathscr{L}}(P_{\mathscr{L}},B_{\mathscr{L}})$.

Now suppose that $B'_{\mathscr{L}} \in \langle\mathbb{B}_{\mathscr{L}}\rangle$ is different from $P^{\dagger}_{\mathscr{L}}$. Then

$$\sup_{P_{\mathscr{L}}\in\mathbb{P}^{*}_{\mathscr{L}}} S^{\log}_{g,\mathscr{L}}(P_{\mathscr{L}},B'_{\mathscr{L}}) \geq S^{\log}_{g,\mathscr{L}}(P^{\dagger}_{\mathscr{L}},B'_{\mathscr{L}}) > S^{\log}_{g,\mathscr{L}}(P^{\dagger}_{\mathscr{L}},P^{\dagger}_{\mathscr{L}}),$$

where the strict inequality follows as seen above. This now shows, that adopting $B'_{\mathscr{L}} \neq P^{\dagger}_{\mathscr{L}}$ leads to an avoidably bad score.

Hence $B_{\mathscr{L}} = P^{\dagger}_{\mathscr{L}}$ is the unique function in $\langle \mathbb{B}_{\mathscr{L}} \rangle$ which minimises $\sup_{P_{\mathscr{L}} \in \mathbb{P}^*_{\mathscr{L}}} S^{\log}_{g,\mathscr{L}}(P_{\mathscr{L}}, B_{\mathscr{L}})$. □

We see, then, that the results of §2 concerning beliefs defined on propositions extend naturally to beliefs defined on the sentences of a propositional language. In light of these findings, our subsequent discussions will, for ease of exposition, solely focus on propositions. It should be clear how our remarks generalise to sentences.

§4
Relationship to standard entropy maximisation

We have seen so far that there is a sense in which our notions of entropy and expected loss depend on the weight given to each partition under consideration—i.e., on the weighting function g. It is natural to demand that no proposition should be entirely dismissed from consideration by being given zero weight—that g be inclusive. In which case, the belief function that minimises worst-case g-expected loss is just the probability function in $\mathbb{E}$ that maximises g-entropy, if there is such a function. This result provides a single justification of the three norms of objective Bayesianism: the belief function should be a probability function, it should be in $\mathbb{E}$, i.e., calibrated to evidence of physical probability, and it should otherwise be equivocal, where the degree to which a belief function is equivocal can be measured by its g-entropy.

This line of argument gives rise to two questions. Which g-entropy should be maximised? Does the standard entropy maximiser count as a rational belief function?

¶ On the former question, the task is to isolate some set $\mathscr{G}$ of appropriate weighting functions. Thus far, the only restriction imposed on a weighting function g has been that it should be inclusive; this is required in order that scoring rules evaluate *all* beliefs, rather than just a select few. We shall put forward two further conditions which can help to narrow down a proper subclass $\mathscr{G}$ of weighting functions.

A second natural desideratum is the following:

Definition 37 (*Symmetric weighting function*). A weighting function g is *symmetric* if and only if whenever π' can be obtained from π by permuting the ω_i in π, then $g(\pi') = g(\pi)$.

For example, for $|\Omega| = 4$ and symmetric g we have that $g(\{\{\omega_1, \omega_2\}, \{\omega_3\}, \{\omega_4\}\}) = g(\{\{\omega_1, \omega_4\}, \{\omega_2\}, \{\omega_3\}\})$. Note that $g_{\Omega}, g_{\mathscr{P}\Omega}$ and g_{Π} are all symmetric. The symmetry condition can also be stated as follows: $g(\pi)$ is only a function of the *spectrum* of π, i.e., of the multi-set of sizes of the members of π. In the above example the spectrum of both partitions is $\{2, 1, 1\}$.

It turns out that inclusive and symmetric weighting functions lead to g-entropy maximisers that satisfy a variety of intuitive and plausible properties—see Appendix B.

In addition, it is natural to suppose that if π' is a refinement of partition π then g should not give any less weight to π' than it does to π—there are no grounds to favour coarser partitions over more fine-grained partitions, although, as Keynes (1921, Chapter 4) argued, there may be grounds to prefer finer-grained partitions over coarser partitions.

Definition 38 (*Refined weighting function*). A weighting function g is *refined* if and only if whenever π' refines π then $g(\pi') \geq g(\pi)$.

g_Π and g_Ω are refined, but $g_{\mathscr{P}\Omega}$ is not.

Let $\mathscr{G}_0$ be the set of weighting functions that are inclusive, symmetric and refined. One might plausibly set $\mathscr{G} = \mathscr{G}_0$. We would at least suggest that all the weighting functions in $\mathscr{G}_0$ are appropriate weighting functions for scoring rules; we shall leave it open as to whether $\mathscr{G}$ should contain some weighting functions—such as the proposition weighting $g_{\mathscr{P}\Omega}$—that lie outside $\mathscr{G}_0$. We shall thus suppose in what follows that the set $\mathscr{G}$ of appropriate weighting functions is such that $\mathscr{G}_0 \subseteq \mathscr{G} \subseteq \mathscr{G}_{\text{inc}}$, where $\mathscr{G}_{\text{inc}}$ is the set of inclusive weighting functions.

¶ One might think that the second question posed above—does the standard entropy maximiser count as a rational belief function?—should be answered in the negative. We saw in §2.2 that the standard entropy, g_Ω-entropy, has a weighting function g_Ω that is not inclusive. So there is no guarantee that the standard entropy maximiser minimises worst-case g-expected loss for some $g \in \mathscr{G}$. Indeed, Fig. 1 showed that the standard entropy maximiser need neither coincide with the partition entropy maximiser nor the proposition entropy maximiser.

However, it would be too hasty to conclude that the standard entropy maximiser fails to qualify as a rational belief function. Recall that the Equivocation norm says that an agent's belief function should be *sufficiently* equivocal, rather than maximally equivocal. This qualification is essential to cope with the situation in which there is no maximally equivocal function in $\mathbb{E}$, i.e., the situation in which for any function in $\mathbb{E}$ there is another function in $\mathbb{E}$ that is more equivocal. This arises, for instance, when one has evidence that a coin is biased in favour of tails, $\mathbb{E} = \mathbb{P}^* = \{P : P(Tails) > 1/2\}$. In this case $\sup_{P\in\mathbb{E}} H_g(P)$ is achieved by the probability function which gives probability 1/2 to tails, which is outside $\mathbb{E}$. This situation also arises in certain cases when evidence is determined by quantified propositions (Williamson, 2013, §2). The best one can do in such a situation is adopt a probability function in $\mathbb{E}$ that is sufficiently equivocal, where what counts as sufficiently equivocal may depend on pragmatic factors such as the required numerical accuracy of predictions and the computational resources available to isolate a suitable function.

Let $\Downarrow\mathbb{E}$ be the set of belief functions that are sufficiently equivocal. Plausibly,[7]

E1: $\Downarrow\mathbb{E} \neq \emptyset$. An agent is always entitled to hold some beliefs.

E2: $\Downarrow\mathbb{E} \subseteq \mathbb{E}$. Sufficiently equivocal belief functions are calibrated with evidence.

E3: For all $g \in \mathscr{G}$ there is some $\epsilon > \inf_{B\in\mathbb{B}} \sup_{P\in\mathbb{E}} S_g^{\log}(P,B)$ such that if $R \in \mathbb{E}$ and $\sup_{P\in\mathbb{E}} S_g(P,R) < \epsilon$ then $R \in \Downarrow\mathbb{E}$. I.e., if R has sufficiently low worst-case g-expected loss for some appropriate g, then R is sufficiently equivocal.

E4: $\Downarrow\Downarrow\mathbb{E} = \Downarrow\mathbb{E}$. Any function, from those that are calibrated with evidence, that is sufficiently equivocal, is a function, from those that are calibrated with evidence and are sufficiently equivocal, that is sufficiently equivocal.

E5: If P is a limit point of $\Downarrow\mathbb{E}$ and $P \in \mathbb{E}$ then $P \in \Downarrow\mathbb{E}$.

[7] A closely related set of conditions was put forward in Williamson (2013). Note that we will not need to appeal to E4 in this paper. E1 is a consequence of the other principles together with the fact that $\mathbb{E} \neq \emptyset$.

Conditions E2, E3 and E5 allow us to answer our two questions. Which g-entropy should be maximised? By E3, it is rational to adopt any g-entropy maximiser that is in $\mathbb{E}$, for $g \in \mathscr{G} \supseteq \mathscr{G}_0$. Does the standard entropy maximiser count as a rational belief function? Yes, if it is in $\mathbb{E}$ (which is the case, for instance, if $\mathbb{E}$ is closed):

Theorem 39 (*Justification of maxent*). *If* $\mathbb{E}$ *contains its standard entropy maximiser,* $P^{\dagger}_{\Omega} := \arg\sup_{\mathbb{E}} H_{\Omega}$, *then* $P^{\dagger}_{\Omega} \in \Downarrow\mathbb{E}$.

Proof: We shall first see that there is a sequence of $(g_t)_{t\in\mathbb{N}}$ in $\mathscr{G}$ such that the g_t-entropy maximisers $P^{\dagger}_t \in [\mathbb{E}]$ converge to $P^{\dagger}_{\Omega}$. All respective entropy maximisers are unique due to Corollary 10.

Let $g_t(\{\{\omega\} : \omega \in \Omega\}) = 1$, and put $g_t(\pi) := \frac{1}{t}$ for all other $\pi \in \Pi$. The g_t are in $\mathscr{G}$ because they are inclusive, symmetric and refined. g_t-entropy has the following form:

$$H_t := \sup_{P\in\mathbb{E}} H_{g_t}(P) = \sup_{P\in\mathbb{E}} \sum_{\pi\in\Pi} -g_t(\pi) \sum_{F\in\pi} P(F)\log P(F).$$

Now note that $g_t(\pi)$ converges to $g_{\Omega}(\pi)$ and that $P(F)\log P(F)$ is finite for all $F \subseteq \Omega$. Thus, for all $P \in \mathbb{P}$ $H_t(P)$ converges to $H_{\Omega}(P)$ as t approaches infinity. Hence, $\sup_{P\in\mathbb{E}} H_{g_t}(P) = H_t$ tends to $\sup_{P\in\mathbb{E}} H_{\Omega}(P) = H_{\Omega}$.

Let us now compute

$$\begin{aligned}
|H_{\Omega}(P^{\dagger}_t) - H_{\Omega}(P^{\dagger}_{\Omega})| &= |H_{\Omega}(P^{\dagger}_t) - H_{g_t}(P^{\dagger}_t) + H_{g_t}(P^{\dagger}_t) - H_{\Omega}(P^{\dagger}_{\Omega})| \\
&\le |H_{\Omega}(P^{\dagger}_t) - H_{g_t}(P^{\dagger}_t)| + |H_{g_t}(P^{\dagger}_t) - H_{\Omega}(P^{\dagger}_{\Omega})| \\
&= |H_{\Omega}(P^{\dagger}_t) - H_{g_t}(P^{\dagger}_t)| + |H_t - H_{\Omega}|.
\end{aligned}$$

As we noted above, g_t converges to g_{Ω}. Furthermore, $(P^{\dagger}_t)_{t\in\mathbb{N}}$ is a bounded sequence. Hence, $H_{g_t}(P^{\dagger}_t)$ converges to $H_{\Omega}(P^{\dagger}_t)$. Also recall that H_t tends to H_{Ω}. Overall, we find that $\lim_{t\to\infty} H_{\Omega}(P^{\dagger}_t) = H_{\Omega}(P^{\dagger}_{\Omega})$.

Since $H_{\Omega}(\cdot)$ is a strictly concave function on $[\mathbb{E}]$ and $[\mathbb{E}]$ is convex, it follows that $P^{\dagger}_t$ converges to $P^{\dagger}_{\Omega}$.

Note that the $P^{\dagger}_t$ are not necessarily in $\mathbb{E}$. But they are in $[\mathbb{E}]$ and there will be some sequence of $P^{\ddagger}_t \in \Downarrow\mathbb{E}$ close to $P^{\dagger}_t$ such that $\lim_{t\to\infty} P^{\ddagger}_t = P^{\dagger}_{\Omega}$, as we shall now see.

If $P^{\dagger}_t \in \mathbb{E}$, then simply let $P^{\ddagger}_t = P^{\dagger}_t$, which is in $\Downarrow\mathbb{E}$ by E3.

If $P^{\dagger}_t \notin \mathbb{E}$, then there exists a $P' \in \mathbb{E}$ which is different from $P^{\dagger}_t$ such that all the points on the line segment between $P^{\dagger}_t$ and P' are in $\mathbb{E}$; with the exception of $P^{\dagger}_t$. Now define $P^{\ddagger}_{t,\delta_t}(\omega) = (1-\delta_t)P^{\dagger}_t(\omega) + \delta_t P'(\omega) = P^{\dagger}_t(\omega) + \delta_t(P'(\omega) - P^{\dagger}_t(\omega))$. Note that for $0 < \delta_t < 1$, we have, for all $\omega \in \Omega$, that $P^{\dagger}_t(\omega) > 0$ implies $P^{\ddagger}_{t,\delta_t}(\omega) > 0$.

Then with

$$m_t := \min_{\substack{\omega\in\Omega \\ P^{\dagger}_t(\omega)>0}} \{P^{\dagger}_t(\omega)\}$$

and $0 < \delta_t < m_t$ it follows from Proposition 70 that for all $F \subseteq \Omega$ and all $P \in \mathbb{E}$, $P(F) > 0$ implies $P^{\dagger}_t(F) > 0$. Thus, for such an F we have $P^{\dagger}_t(F) \ge m_t > \delta_t > 0$.

We find for $P \in [\mathbb{E}]$ and $m_t > \delta_t$ that,[8]

$$\begin{aligned}
|S^{\log}_{g_t}(P,P^{\ddagger}_{t,\delta_t}) - S^{\log}_{g_t}(P,P^{\dagger}_t)| &\leq \sum_{\pi\in\Pi} g_t(\pi) | \sum_{F\in\pi} P(F)\Big(\log P^{\ddagger}_{t,\delta_t}(F) - \log P^{\dagger}_t(F)\Big)| \\
&\leq \sum_{\pi\in\Pi} g_t(\pi) \sum_{\substack{F\in\pi \\ P(F)>0}} P(F) | \log P^{\ddagger}_{t,\delta_t}(F) - \log P^{\dagger}_t(F)| \\
&\leq \sum_{\pi\in\Pi} g_t(\pi) \sum_{\substack{F\in\pi \\ P(F)>0}} P(F) | \log \frac{P^{\dagger}_t(F) - \delta_t \cdot |P'(F) - P^{\dagger}_t(F)|}{P^{\dagger}_t(F)} | \\
&\leq \sum_{\pi\in\Pi} g_t(\pi) \sum_{\substack{F\in\pi \\ P(F)>0}} P(F) | \log \frac{P^{\dagger}_t(F) - \delta_t}{P^{\dagger}_t(F)} | \\
&\leq \sum_{\pi\in\Pi} g_t(\pi) \sum_{\substack{F\in\pi \\ P(F)>0}} P(F) | \log \frac{m_t - \delta_t}{m_t} | \\
&= | \log \frac{m_t - \delta_t}{m_t} | \sum_{\pi\in\Pi} g_t(\pi).
\end{aligned}$$

For fixed g_t and all $P \in [\mathbb{E}]$, $|S^{\log}_{g_t}(P,P^{\ddagger}_{t,\delta_t}) - S^{\log}_{g_t}(P,P^{\dagger}_t)|$ becomes arbitrarily small for small δ_t, moreover the upper bound we established does not depend on P. In particular, for all $\chi_t > 0$ there exists a $T \in \mathbb{N}$ such that for all $U_t > T$ and all $P \in [\mathbb{E}]$ it holds that $|S^{\log}_{g_t}(P,P^{\ddagger}_{t,\frac{1}{U_t}}) - S^{\log}_{g_t}(P,P^{\dagger}_t)| < \chi_t$.

Now let $\epsilon_t > \inf_{B\in\mathbb{B}} \sup_{P\in\mathbb{E}} S^{\log}_{g_t}(P,B) = H_t$. Then with $\chi_t = \frac{\epsilon_t - H_t}{2} > 0$ we have for big enough U_t that

$$\sup_{P\in\mathbb{E}} S^{\log}_{g_t}(P,P^{\ddagger}_{t,\frac{1}{U_t}}) - \sup_{P\in\mathbb{E}} S^{\log}_{g_t}(P,P^{\dagger}_t) \leq \chi_t.$$

Thus,

$$\begin{aligned}
\sup_{P\in\mathbb{E}} S^{\log}_{g_t}(P,P^{\ddagger}_{t,\frac{1}{U_t}}) &\leq \chi_t + \sup_{P\in\mathbb{E}} S^{\log}_{g_t}(P,P^{\dagger}_t) \\
&= \frac{\epsilon_t - H_t}{2} + H_t \\
&< \epsilon_t.
\end{aligned}$$

Hence, $P^{\ddagger}_{t,\delta_t} \in \Downarrow\mathbb{E}$ by E3 for small enough δ_t. since the worst-case g_t-expected loss of $P^{\ddagger}_{t,\delta_t}$ becomes arbitrarily close to H_t.

Now pick a sequence $\delta_t \searrow 0$ such that δ_t is small enough to ensure that for every t it holds that $P^{\ddagger}_{t,\delta_t} \in \Downarrow\mathbb{E}$. Clearly, the sequence $(P^{\ddagger}_{t,\delta_t})_{t\in\mathbb{N}}$ converges to the limit of the sequence $P^{\dagger}_t$, and this limit is $P^{\dagger}_{\Omega}$. So, the sequence $P^{\ddagger}_{t,\delta_t}$ converges to $P^{\dagger}_{\Omega}$ which is, by our assumption, in $\mathbb{E}$.

By E5 we have $P^{\dagger}_{\Omega} \in \Downarrow\mathbb{E}$. □

So far we have seen that, as long as the standard entropy maximiser is not ruled out by the available evidence, it is sufficiently equivocal and hence it is rational

[8] We shall make the purely notational but very helpful convention that $0(\log 0 - \log 0) = 0$.

for an agent to adopt this function as her belief function. On the other hand, the above considerations also imply that if the entropy maximiser $P^{\dagger}_{\Omega}$ is ruled out by the available evidence (i.e., $P^{\dagger}_{\Omega} \in [\mathbb{E}] \setminus \mathbb{E}$), it is rational to adopt some function P close enough to $P^{\dagger}_{\Omega}$, because such a function will be sufficiently equivocal:

Corollary 40. *For all $\epsilon > 0$ there exists a $P \in \Downarrow\mathbb{E}$ such that $|P(\omega) - P^{\dagger}_{\Omega}(\omega)| < \epsilon$ for all $\omega \in \Omega$.*

Proof: Consider the same sequence g_t as in the above proof. Recall that $P^{\dagger}_{t}$ converges to $P^{\dagger}_{\Omega}$. Now pick a t such that $|P^{\dagger}_{t}(\omega) - P^{\dagger}_{\Omega}(\omega)| < \frac{\epsilon}{2}$ for all $\omega \in \Omega$. For this t it holds that $P^{\ddagger}_{t,\delta_t} \in \Downarrow\mathbb{E}$ for small enough δ_t and that $P^{\ddagger}_{t,\delta_t}$ converges to $P^{\dagger}_{t}$. Thus, for small enough δ_t we have $|P^{\dagger}_{t}(\omega) - P^{\ddagger}_{t,\delta_t}(\omega)| < \frac{\epsilon}{2}$ for all $\omega \in \Omega$. Thus, $|P^{\ddagger}_{t,\delta_t}(\omega) - P^{\dagger}_{\Omega}(\omega)| < \epsilon$ for all $\omega \in \Omega$. □

¶ Is there anything that makes the standard entropy maximiser stand out among all those functions that are sufficiently equivocal? One consideration is language invariance. Suppose $g^{\mathscr{L}}$ is a family of weighting functions, defined for each $\mathscr{L}$. $g^{\mathscr{L}}$ is *language invariant* as long as merely adding new propositional variables to the language does not undermine the $g^{\mathscr{L}}$-entropy maximiser:

Definition 41 (*Language invariant family of weighting functions*). Suppose we are given as usual a set $\mathbb{E}$ of probability functions on a fixed language $\mathscr{L}$. For any $\mathscr{L}'$ extending $\mathscr{L}$, let $\mathbb{E}' = \mathbb{E} \times \mathbb{P}_{\mathscr{L}' \setminus \mathscr{L}}$ be the translation of $\mathbb{E}$ into the richer language $\mathscr{L}'$. A family of weighting functions is *language invariant* if for any such $\mathbb{E}, \mathscr{L}$, any $P^{\dagger} \in \arg\sup_{P \in \mathbb{E}} H_{g^{\mathscr{L}}}(P)$ on $\mathscr{L}$ and any language $\mathscr{L}'$ extending $\mathscr{L}$, there is some $P^{\ddagger} \in \arg\sup_{P \in \mathbb{E}'} H_{g^{\mathscr{L}'}}(P)$ on $\mathscr{L}'$ such that $P^{\ddagger}_{\restriction \mathscr{L}} = P^{\dagger}$, i.e., $P^{\ddagger}(\omega) = P^{\dagger}(\omega)$ for each state ω of $\mathscr{L}$.

It turns out that many families of weighting functions—including the partition weightings and the proposition weightings—are *not* language invariant:

Proposition 42. *The family of partition weightings g_{Π} and the family of proposition weightings $g_{\mathscr{P}\Omega}$ are not language invariant.*

Proof: Let $\mathscr{L} = \{A_1, A_2\}$ and $\mathbb{E} = \{P \in \mathbb{P} : P(\omega_1) + 2P(\omega_2) + 3P(\omega_3) + 4P(\omega_4) = 1.7\}$. The partition entropy maximiser $P^{\dagger}_{\Pi}$ and the proposition entropy maximiser $P^{\dagger}_{\mathscr{P}\Omega}$ for this language and this set $\mathbb{E}$ of calibrated functions are given in the first two rows of the table below.

	ω_1	ω_2	ω_3	ω_4
$P^{\dagger}_{\Pi}$	0.5331	0.2841	0.1324	0.0504
$P^{\dagger}_{\mathscr{P}\Omega}$	0.5192	0.3008	0.1408	0.0392

	χ_1	χ_2	χ_3	χ_4	χ_5	χ_6	χ_7	χ_8
$P^{\ddagger}_{\Pi}$	0.2649	0.2649	0.1441	0.1441	0.0671	0.0671	0.0239	0.0239
$P^{\ddagger}_{\mathscr{P}\Omega}$	0.2510	0.2510	0.1594	0.1594	0.0783	0.0783	0.0113	0.0113

We now add one propositional variable, A_3, to $\mathscr{L}$ and, thus, obtain $\mathscr{L}'$. Denote the states of $\mathscr{L}'$ by $\chi_1 = \omega_1 \wedge \neg A_3$, $\chi_2 = \omega_1 \wedge A_3$, and so on. Assuming that we have no information at all concerning A_3, the set of calibrated probability functions is given by the solutions of the constraint, $(P'(\chi_1)+P'(\chi_2))+2(P'(\chi_3)+P'(\chi_4))+3(P'(\chi_5)+P'(\chi_6))+4(P'(\chi_7)+P'(\chi_8))=1.7$. Language invariance would now entail that $P^\dagger(\omega_1)=P^\ddagger(\chi_1)+P^\ddagger(\chi_2)$, $P^\dagger(\omega_2)=P^\ddagger(\chi_3)+P^\ddagger(\chi_4)$, $P^\dagger(\omega_3)=P^\ddagger(\chi_5)+P^\ddagger(\chi_6)$, $P^\dagger(\omega_4)=P^\ddagger(\chi_7)+P^\ddagger(\chi_8)$. However, neither the partition entropy maximisers nor the proposition entropy maximisers form a language invariant family, as can be seen from the last two rows of the above table. □

On the other hand, it is well known that standard entropy maximisation is language invariant (p. 76 in (Paris, 1994)). This can be seen to follow from the fact that certain families of weighting functions that only assign positive weight to a single partition are language invariant:

Lemma 43. *Suppose a function f picks out a partition π for any language $\mathscr{L}$, in such a way that if $\mathscr{L}'\supseteq\mathscr{L}$ then $f(\mathscr{L}')$ is a refinement of $f(\mathscr{L})$, with each $F\in f(\mathscr{L})$ being refined into the same number k of members $F_1,\dots,F_k\in f(\mathscr{L}')$, for $k\geq 1$. Suppose $g^{\mathscr{L}}$ is such that for any $\mathscr{L}$, $g^{\mathscr{L}}(f(\mathscr{L}))=c>0$ but $g^{\mathscr{L}}(\pi)=0$ for all other partitions π. Then $g^{\mathscr{L}}$ is language invariant.*

Proof: Let $P^\dagger$ denote a $g^{\mathscr{L}}$-entropy maximiser (in $[\mathbb{E}]$), and let $P^\ddagger$ denote a $g^{\mathscr{L}'}$-entropy maximiser in $[\mathbb{E}]\times\mathbb{P}_{\mathscr{L}'\setminus\mathscr{L}}$. Since $g^{\mathscr{L}}$ and $g^{\mathscr{L}'}$ need not be inclusive, $H_{g,\mathscr{L}}$ and $H_{g,\mathscr{L}'}$ need not be strictly concave. Thus, there need not be unique entropy maximisers. Given $F\subseteq\Omega$ refined into subsets $F_1,\dots,F_k$ of Ω', $F'\subseteq\Omega'$ is defined by $F':=F_1\cup\ldots\cup F_k$. One can restrict $P^\ddagger$ to $\mathscr{L}$ by setting $P^\ddagger(\omega)=\sum_{\omega'\in\Omega',\omega'\models\omega}P^\ddagger(\omega')$ for $\omega\in\Omega$, so in particular, $P^\ddagger(F)=P^\ddagger(F')=P^\ddagger(F_1)+\ldots+P^\ddagger(F_k)$ for $F\in\Omega$.

The $g^{\mathscr{L}}$-entropy of $P^\dagger$ is closely related to the $g^{\mathscr{L}'}$-entropy of $P^\ddagger$:

$$\begin{aligned}
-c\sum_{F\in f(\mathscr{L})}&P^\dagger(F)\log P^\dagger(F)\\
&\geq -c\sum_{F\in f(\mathscr{L})}P^\ddagger(F)\log P^\ddagger(F)\\
&=-c\sum_{F\in f(\mathscr{L})}(P^\ddagger(F_1)+\ldots+P^\ddagger(F_k))\log(P^\ddagger(F_1)+\ldots+P^\ddagger(F_k))\\
&=-c\sum_{F\in f(\mathscr{L})}(P^\ddagger(F_1)+\ldots+P^\ddagger(F_k))\Big(\log k+\log\frac{P^\ddagger(F_1)+\ldots+P^\ddagger(F_k)}{k}\Big)\\
&\overset{LSI}{\geq}-c\log k-c\sum_{F\in f(\mathscr{L})}P^\ddagger(F_1)\log P^\ddagger(F_1)+\ldots+P^\ddagger(F_k)\log P^\ddagger(F_k)\\
&=-c\log k-c\sum_{G\in f(\mathscr{L}')}P^\ddagger(G)\log P^\ddagger(G)\\
&=-c\log k-c\sum_{F\in f(\mathscr{L})}P^\ddagger(F_1)\log P^\ddagger(F_1)+\ldots+P^\ddagger(F_k)\log P^\ddagger(F_k)\\
&\geq -c\log k-c\sum_{F\in f(\mathscr{L})}\frac{P^\dagger(F)}{k}\log\frac{P^\dagger(F)}{k}+\ldots+\frac{P^\dagger(F)}{k}\log\frac{P^\dagger(F)}{k}\\
&=-c\log k-c\sum_{F\in f(\mathscr{L})}P^\dagger(F)\log\frac{P^\dagger(F)}{k}\\
&=-c\sum_{F\in f(\mathscr{L})}P^\dagger(F)\log P^\dagger(F).
\end{aligned}$$

LSI refers to the log sum inequality introduced in Lemma 16. The first and last inequality above follow from the fact that $P^\dagger$ and $P^\ddagger$ are entropy maximisers over $\mathscr{L}$, $\mathscr{L}'$ respectively. Hence, all inequalities are indeed equalities. These entropy maximisers are unique on $f(\mathscr{L}), f(\mathscr{L}')$, so $P^\dagger(F) = k \cdot P^\ddagger(F_1) = \ldots = k \cdot P^\ddagger(F_k) = P^\ddagger(F)$ for $F \in f(\mathscr{L})$.

Now take an arbitrary $P^\dagger \in \arg\sup_{P \in \mathbb{E}} H_{g^{\mathscr{L}}}(P)$ and suppose $\omega \in \Omega$. Any $P^\ddagger$ such that $P^\ddagger(\omega) = P^\dagger(\omega)$ and $P^\ddagger(F_1) = \ldots = P^\ddagger(F_k) = P^\dagger(F)/k$ will be a $g^{\mathscr{L}'}$-entropy maximiser on $\mathscr{L}'$. Thus $g^{\mathscr{L}}$ is language invariant.

Note that if, for some $\mathscr{L}$, $f(\mathscr{L}) = \{\Omega^{\mathscr{L}}, \emptyset\}$, where $\Omega^{\mathscr{L}}$ denotes the set of states of $\mathscr{L}$, then $H_{g^{\mathscr{L}}}(P) = -P(\Omega^{\mathscr{L}})\log P(\Omega^{\mathscr{L}}) - P(\emptyset)\log P(\emptyset) = 0 - 0 = 0$. Likewise, if $f(\mathscr{L}') = \{\Omega^{\mathscr{L}'}\}$, then $H_{g^{\mathscr{L}'}}(P) = 0$. For such g-entropies, every probability maximises g-entropy trivially since all probability functions have the same g-entropy. □

Taking $f(\mathscr{L}) = \{\{\omega\} : \omega \in \Omega\}$ and $c = 1$ we have the language invariance of standard entropy maximisation:

Corollary 44. *The family of weighting functions g_Ω is language invariant.*

While giving weight in this way to just one partition is sufficient for language invariance, it is not necessary, as we shall now see. Define a family of weighting functions, the *substate* weighting functions, by giving weight to just those partitions that are partitions of states of sublanguages. For any sublanguage $\mathscr{L}^- \subseteq \mathscr{L} = \{A_1, \ldots, A_n\}$, let Ω^- be the set of states of $\mathscr{L}^-$ and let π^- be the partition of propositions of $\mathscr{L}$ that represents the partition of states of the sublanguage $\mathscr{L}^-$, i.e., $\pi^- = \{\{\omega \in \Omega : \omega \models \omega^-\} : \omega^- \in \Omega^-\}$. Then,

$$g_{\subseteq}^{\mathscr{L}}(\pi) = \begin{cases} 1 & : \quad \pi = \pi^- \text{ for some } \mathscr{L}^- \subseteq \mathscr{L} \\ 0 & : \quad \text{otherwise} \end{cases} .$$

Example 45. For $\mathscr{L} = \{A_1, A_2\}$ there are three sublanguages: $\mathscr{L}$ itself and the two proper sublanguages: $\{A_1\}, \{A_2\}$. Then $g_{\subseteq}^{\mathscr{L}}$ assigns the following three partitions of Ω the same positive weight: $\{\{A_1 \wedge A_2, A_1 \wedge \neg A_2\}, \{\neg A_1 \wedge A_2, \neg A_1 \wedge \neg A_2\}\}$, $\{\{A_1 \wedge A_2, \neg A_1 \wedge A_2\}, \{A_1 \wedge \neg A_2, \neg A_1 \wedge \neg A_2\}\}$, $\{\{A_1 \wedge A_2\}, \{A_1 \wedge \neg A_2\}, \{\neg A_1 \wedge A_2\}, \{\neg A_1 \wedge \neg A_2\}\}$. $g_{\subseteq}^{\mathscr{L}}$ assigns all other $\pi \in \Pi$ weight zero.

Note that there are $2^n - 1$ non-empty sublanguages of $\mathscr{L}$, so $g_{\subseteq}^{\mathscr{L}}$ gives positive weight to $2^n - 1$ partitions.

Proposition 46. *The family of substate weighting functions is language invariant.*

Proof: Consider an extension $\mathscr{L}' = \{A_1, \ldots, A_n, A_{n+1}\}$ of $\mathscr{L}$. Let $P^\dagger, P^\ddagger$ be $g_\subseteq$-entropy maximisers on $\mathscr{L}, \mathscr{L}'$ respectively. For simplicity of exposition we shall view these functions as defined over sentences so that we can talk of $P^\ddagger(A_{n+1} \wedge \omega^-)$ etc. For the purposes of the following calculation we shall consider the empty language to be a language. Entropies over the empty language vanish. Summing over the empty language ensures, for example, that the expression $P^\ddagger(A_{n+1})\log P^\ddagger(A_{n+1})$ appears in Equation 27.

$$
\begin{aligned}
2H_{g_{\subseteq}^{\mathscr{L}}}(P^{\dagger}) &= -2\sum_{\mathscr{L}^{-}\subseteq\mathscr{L}}\sum_{\omega^{-}\in\Omega^{-}} P^{\dagger}(\omega^{-})\log P^{\dagger}(\omega^{-}) \\
&\geq -2\sum_{\mathscr{L}^{-}\subseteq\mathscr{L}}\sum_{\omega^{-}\in\Omega^{-}} P^{\ddagger}(\omega^{-})\log P^{\ddagger}(\omega^{-}) \\
&= -\sum_{\mathscr{L}^{-}\subseteq\mathscr{L}}\sum_{\omega^{-}\in\Omega^{-}} P^{\ddagger}(\omega^{-})\log P^{\ddagger}(\omega^{-}) \\
&\quad -\sum_{\mathscr{L}^{-}\subseteq\mathscr{L}}\sum_{\omega^{-}\in\Omega^{-}} \left[P^{\ddagger}(A_{n+1}\wedge\omega^{-})+P^{\ddagger}(\neg A_{n+1}\wedge\omega^{-})\right] \\
&\qquad\qquad \times\log\left[P^{\ddagger}(A_{n+1}\wedge\omega^{-})+P^{\ddagger}(\neg A_{n+1}\wedge\omega^{-})\right] \\
&= -\sum_{\mathscr{L}^{-}\subseteq\mathscr{L}}\sum_{\omega^{-}\in\Omega^{-}} P^{\ddagger}(\omega^{-})\log P^{\ddagger}(\omega^{-}) \\
&\quad -\sum_{\mathscr{L}^{-}\subseteq\mathscr{L}}\sum_{\omega^{-}\in\Omega^{-}} \left[P^{\ddagger}(A_{n+1}\wedge\omega^{-})+P^{\ddagger}(\neg A_{n+1}\wedge\omega^{-})\right] \\
&\qquad\qquad \times\log\left[2\cdot\frac{P^{\ddagger}(A_{n+1}\wedge\omega^{-})+P^{\ddagger}(\neg A_{n+1}\wedge\omega^{-})}{1+1}\right] \\
&\geq -\sum_{\mathscr{L}^{-}\subseteq\mathscr{L}}\sum_{\omega^{-}\in\Omega^{-}} P^{\ddagger}(\omega^{-})\log P^{\ddagger}(\omega^{-}) \\
&\quad -\sum_{\mathscr{L}^{-}\subseteq\mathscr{L}}\sum_{\omega^{-}\in\Omega^{-}} [\log 2+P^{\ddagger}(A_{n+1}\wedge\omega^{-})\log P^{\ddagger}(A_{n+1}\wedge\omega^{-}) \\
&\quad +P^{\ddagger}(\neg A_{n+1}\wedge\omega^{-})\log P^{\ddagger}(\neg A_{n+1}\wedge\omega^{-})] \\
&= -c\log 2-\sum_{\substack{\mathscr{L}^{-}\subseteq\mathscr{L}'\\ \{A_{n+1}\}\notin\mathscr{L}'}}\sum_{\omega^{-}\in\Omega^{-}} P^{\ddagger}(\omega^{-})\log P^{\ddagger}(\omega^{-}) \\
&\quad -\sum_{\substack{\mathscr{L}^{-}\subseteq\mathscr{L}'\\ \{A_{n+1}\}\in\mathscr{L}'}}\sum_{\omega^{-}\in\Omega^{-}} P^{\ddagger}(\omega^{-})\log P^{\ddagger}(\omega^{-}) \\
&= -c\log 2-\sum_{\mathscr{L}^{-}\subseteq\mathscr{L}'}\sum_{\omega^{-}\in\Omega^{-}} P^{\ddagger}(\omega^{-})\log P^{\ddagger}(\omega^{-}) \\
&= -c\log 2+H_{g_{\subseteq}^{\mathscr{L}'}}(P^{\ddagger}) \\
&= -c\log 2-\sum_{\mathscr{L}^{-}\subseteq\mathscr{L}}\sum_{\omega^{-}\in\Omega^{-}} P^{\ddagger}(\omega^{-})\log P^{\ddagger}(\omega^{-}) \\
&\quad -\sum_{\mathscr{L}^{-}\subseteq\mathscr{L}}\sum_{\omega^{-}\in\Omega^{-}} [P^{\ddagger}(A_{n+1}\wedge\omega^{-})\log P^{\ddagger}(A_{n+1}\wedge\omega^{-}) \\
&\quad +P^{\ddagger}(\neg A_{n+1}\wedge\omega^{-})\log P^{\ddagger}(\neg A_{n+1}\wedge\omega^{-})] \\
&\geq -c\log 2-\sum_{\mathscr{L}^{-}\subseteq\mathscr{L}}\sum_{\omega^{-}\in\Omega^{-}} P^{\dagger}(\omega^{-})\log P^{\dagger}(\omega^{-})-\sum_{\mathscr{L}^{-}\subseteq\mathscr{L}}\sum_{\omega^{-}\in\Omega^{-}} P^{\dagger}(\omega^{-})\log\frac{P^{\dagger}(\omega^{-})}{2} \\
&= -2\sum_{\mathscr{L}^{-}\subseteq\mathscr{L}}\sum_{\omega^{-}\in\Omega^{-}} P^{\dagger}(\omega^{-})\log P^{\dagger}(\omega^{-}) \\
&= 2H_{g_{\subseteq}^{\mathscr{L}}}(P^{\dagger}),
\end{aligned}
\tag{27}
$$

where c is some constant and where the second inequality is an application of the log-sum inequality. As in the previous proof, all inequalities are thus equalities, $P^{\ddagger}(\pm A_{n+1}\wedge\omega)=P^{\dagger}(\omega)/2$ and $P^{\ddagger}$ extends $P^{\dagger}$, as required. □

In general the substate entropy maximisers differ from the standard entropy maximisers as well as the partition entropy maximisers and the proposition entropy maximisers:

Example 47. For $\mathscr{L} = \{A_1, A_2\}$ and the substate weighting function $g^{\mathscr{L}}_{\subseteq}$ on $\mathscr{L}$ (see Example 45) we find for $\mathbb{E} = \{P \in \mathbb{P} : P(A_1 \wedge A_2) + 2P(A_1 \wedge \neg A_2) = 0.1\}$ that the standard entropy maximiser, the partition entropy maximiser, the proposition entropy maximiser and the substate weighting entropy maximiser are pairwise different.

	$A_1 \wedge A_2$	$A_1 \wedge \neg A_2$	$\neg A_1 \wedge A_2$	$\neg A_1 \wedge \neg A_2$
$P^{\dagger}_{\Omega}$	0.0752	0.0124	0.4562	0.4562
$P^{\dagger}_{\Pi}$	0.0856	0.0072	0.4536	0.4536
$P^{\dagger}_{\mathscr{P}\Omega}$	0.0950	0.0025	0.4513	0.4513
$P^{\dagger}_{g^{\mathscr{L}}_{\subseteq}}$	0.0950	0.0025	0.4293	0.4732

Observe that the standard entropy maximiser, the partition entropy maximiser and the proposition entropy maximiser are all symmetric in $\neg A_1 \wedge A_2$ and $\neg A_1 \wedge \neg A_2$, while the substate weighting entropy maximiser is not. This break of symmetry is caused by the fact that $g^{\mathscr{L}}_{\subseteq}$ is not symmetric in $\neg A_1 \wedge A_2$ and $\neg A_1 \wedge \neg A_2$.

We have seen that the substate weighting functions are not symmetric. Neither are they inclusive nor refined. We conjecture that, if $\mathscr{G} = \mathscr{G}_0$, the set of inclusive, symmetric and refined g, then the only language invariant family $g^{\mathscr{L}}$ that gives rise to entropy maximisers that are sufficiently equivocal is the family that underwrites standard entropy maximisation: if $g^{\mathscr{L}}$ is language invariant and the $g^{\mathscr{L}}$-entropy maximiser is in $\Downarrow\mathbb{E}$ then $g^{\mathscr{L}} = g_{\Omega}$.

In sum, there is a compelling reason prefer the standard entropy maximiser over other g-entropy maximisers: the standard entropy maximiser is language invariant while other—perhaps, all other—appropriate g-entropy maximisers are not. In Appendix B.3 we show that there are three further ways in which the standard entropy maximiser differs from other g-entropy maximisers: it satisfies the principles of Irrelevance, Relativisation, and Independence.

§5
Discussion

§5.1. Summary

In this paper we have seen how the standard concept of entropy generalises rather naturally to the notion of g-entropy, where g is a function that weights the partitions that contribute to the entropy sum. If loss is taken to be logarithmic, as is forced by desiderata L1–4 for a default loss function, then the belief function that minimises worst-case g-expected loss, where the expectation is taken with respect to a chance function known to lie in a convex set $\mathbb{E}$, is the probability function in $\mathbb{E}$ that maximises g-entropy, if there is such a function. This applies whether belief functions are thought of as defined over the sentences of an agent's language or over the propositions picked out by those sentences.

This fact suggests a justification of the three norms of objective Bayesianism: a belief function should be a probability function, it should lie in the set $\mathbb{E}$ of potential chance functions, and it should otherwise be equivocal in that it should have maximum g-entropy.

But the probability function with maximum g-entropy may lie outside $\mathbb{E}$, on its boundary, in which case that function is ruled out of contention by available evidence. So objective Bayesianism only requires that a belief function be *sufficiently* equivocal—not that it be maximally equivocal. Principles E1–5 can be used to constrain the set $\Downarrow\mathbb{E}$ of sufficiently equivocal functions. Arguably, if the standard entropy maximiser is in $\mathbb{E}$ then it is also in $\Downarrow\mathbb{E}$. Moreover, the standard entropy maximiser stands out as being language invariant. This then provides a qualified justification of the standard maximum entropy principle: while an agent is rationally entitled to adopt any sufficiently equivocal probability function in $\mathbb{E}$ as her belief function, if the standard entropy maximiser is in $\mathbb{E}$ then that function is a natural choice.

Some questions arise. First, what are the consequences of this sort of account for conditionalisation and Bayes' theorem? Second, how does this account relate to imprecise probability, advocates of which reject our starting assumption that the strengths of an agent's beliefs are representable by a single belief function? Third, the arguments of this paper are overtly pragmatic; can they be reformulated in a non-pragmatic way? We shall tackle these questions in turn.

§5.2. Conditionalisation, conditional probabilities and Bayes' theorem

Subjective Bayesians endorse the Probability norm and often also some sort of Calibration norm, but do not go so far as to insist on Equivocation. This leads to relatively weak constraints on degrees of belief, so subjective Bayesians typically appeal to Bayesian conditionalisation as a means to tightly constrain the way in which degrees of belief change in the light of new evidence. Objective Bayesians do not need to invoke Bayesian conditionalisation as a norm of belief change because the three norms of objective Bayesianism already tightly constrain any new belief function that an agent can adopt. In fact, if the objective Bayesian adopts the policy of adopting the standard entropy maximiser as her belief function then objective Bayesian updating often agrees with updating by conditionalisation, as shown by Seidenfeld (1986, Result 1):

Theorem 48. *Suppose that $\mathbb{E}$ is the set of probability functions calibrated with evidence E and that $\mathbb{E}$ can be written as the set of probability functions which satisfy finitely many constraints of the form $c_i = \sum_{\omega\in\Omega} d_{i,\omega}P(\omega)$. Suppose $\mathbb{E}'$ is the set of probability functions calibrated with evidence $E\cup\{G\}$, and that $P^\dagger_E, P^\dagger_{E\cup\{G\}}$ are functions in $\mathbb{E}, \mathbb{E}'$ respectively that maximise standard entropy. If*

(i) $G\subseteq\Omega$,

(ii) the only constraints imposed by $E\cup\{G\}$ are the constraints $c_i = \sum_{\omega\in\Omega} d_{i,\omega}P(\omega)$ imposed by E together with the constraint $P(G)=1$,

(iii) the constraints in (ii) are consistent, and

(iv) $P^\dagger_E(\cdot|G)\in\mathbb{E}$,

then $P^\dagger_{E\cup\{G\}}(F) = P^\dagger_E(F|G)$ for all $F\subseteq\Omega$.

This fact has various consequences. First, it provides a qualified justification of Bayesian conditionalisation: a standard entropy maximiser can be thought of as applying Bayesian conditionalisation in many natural situations. Second, if conditions (i)-(iv) of Theorem 48 hold then there is no need to maximise standard entropy to compute the agent's new degrees of belief—instead, Bayesian conditionalisation can be used to calculate these degrees of belief. Third, conditions (i)-(iv) of Theorem 48 can each fail, so the two forms of updating do not always agree and Bayesian conditionalisation is less central to an objective Bayesian who maximises standard entropy than it is to a subjective Bayesian. As pointed out in Williamson (2010, Chapter 4) and Williamson (2011, §§8,9), standard entropy maximisation is to be preferred over Bayesian conditionalisation where any of these conditions fail. Fourth, conditional probabilities, which are crucial to subjective Bayesianism on account of their use in Bayesian conditionalisation, are less central to the objective Bayesian, because conditionalisation is only employed in a qualified way. For the objective Bayesian, conditional probabilities are merely ratios of unconditional probabilities—they are not generally interpretable as conditional degrees of belief (Williamson, 2010, §4.4.1). Fifth, Bayes' theorem, which is an important tool for calculating conditional probabilities, used routinely in Bayesian statistics, for example, is less central to objective Bayesianism because of the less significant role played by conditional probabilities.

Interestingly, while Theorem 48 appeals to standard entropy maximisation, an analogous result holds for g-entropy maximisation, for any inclusive g, as we show in Appendix B.2:

Theorem 49. *Suppose that convex and closed $\mathbb{E}$ is the set of probability functions calibrated with evidence E, and $\mathbb{E}'$ is the set of probability functions calibrated with evidence $E \cup \{G\}$. Also suppose that $P^{\dagger}_{E}, P^{\dagger}_{E\cup\{G\}}$ are functions in $\mathbb{E}, \mathbb{E}'$ respectively that maximise g-entropy for some fixed $g \in \mathscr{G}_{\text{inc}} \cup \{g_{\Omega}\}$. If*

(i) $G \subseteq \Omega$,

(ii) the only constraints imposed by $E \cup \{G\}$ are the constraints imposed by E together with the constraint $P(G) = 1$,

(iii) the constraints in (ii) are consistent, and

(iv) $P^{\dagger}_{E}(\cdot|G) \in \mathbb{E}$,

then $P^{\dagger}_{E\cup\{G\}}(F) = P^{\dagger}_{E}(F|G)$ for all $F \subseteq \Omega$.

Thus the preceding comments apply equally in the more general context of this paper.

§5.3. Imprecise probability

Advocates of imprecise probability argue that an agent's belief state is better represented by a set of probability functions—for example by the set $\mathbb{E}$ of probability functions calibrated with evidence—than by a single belief function (Kyburg Jr, 2003). This makes decision making harder. An agent whose degrees of belief are represented by a single probability function can use that probability function to determine which of the available acts maximises expected utility. However, an imprecise agent will typically find that the acts that maximise expected utility vary according to which probability function in her imprecise belief state is used to determine the expectation. The question then arises, with respect to which probability function in her belief state should such expectations be taken?

This question motivates a two-step procedure for imprecise probability: first isolate a set of probability functions as one's belief state; then choose a probability function from within this set for decision making—this might be done in advance of any particular decision problem arising—, and use that function to make decisions by maximising expected utility. While this sort of procedure is not the only way of thinking about imprecise probability, it does have some adherents. It is a component of the transferrable belief model of Smets and Kennes (1994), for instance, and Keynes advocated a similar sort of view:[9]

> the prospect of a European war is uncertain, or the price of copper and the rate of interest twenty years hence, or the obsolescence of a

[9] We are very grateful to an anonymous referee and Hykel Hosni respectively for alerting us to these two views.

> new invention, or the position of private wealth-owners in the social system in 1970. About these matters there is no scientific basis on which to form any calculable probability whatever. We simply do not know. Nevertheless, the necessity for action and for decision compels us as practical men to do our best to overlook this awkward fact and to behave exactly as we should if we had behind us a good Benthamite calculation of a series of prospective advantages and disadvantages, each multiplied by its appropriate probability, waiting to be summed. (Keynes, 1937, p.214.)

The results of this paper can be applied at the second step of this two-step procedure. If one wants a probability function for decision making that controls worst-case g-expected default loss, then one should choose a function in one's belief state with sufficiently high g-entropy (or a limit point of such functions), where g is in $\mathscr{G}$, the set of appropriate weighting functions. The resulting approach to imprecise probability is conceptually different to objective Bayesian epistemology, but the two approaches are formally equivalent, with the decision function for imprecise probability corresponding to the belief function for objective Bayesian epistemology.

§5.4. A non-pragmatic justification

The line of argument in this paper is thoroughly pragmatic: one ought to satisfy the norms of objective Bayesianism in order to control worst-case expected *loss*. However, the question has recently arisen as to whether one can adapt arguments that appeal to scoring rules to provide a non-pragmatic justification of the norms of rational belief—see, e.g., Joyce (2009). There appears to be some scope for reinterpreting the arguments of this paper in non-pragmatic terms, along the following lines. Instead of viewing L1–4 as isolating an appropriate default loss function, one can view them as postulates on a measure of the inaccuracy of one's belief in a true proposition: believing a true proposition does not expose one to inaccuracy; inaccuracy strictly increases as degree of belief in the true proposition decreases; inaccuracy with respect to a proposition only depends on the degree of belief in that proposition; inaccuracy is additive over independent sublanguages.[10] A g-scoring rule then measures expected inaccuracy. Strict propriety implies that the physical probability function has minimum expected inaccuracy. (If P^* is deterministic, i.e., $P^*(\omega) = 1$ for some $\omega \in \Omega$, then the unique probability function which puts all mass on ω has minimum expected inaccuracy. In this sense we can say that strictly proper scoring rules are truth-tracking, which is an important epistemic good.) In order to minimise worst-case g-expected inaccuracy, one would need degrees of belief that are probabilities, that are calibrated to phyisical probability, and that maximise g-entropy.

The main difference between the pragmatic and the non-pragmatic interpretations of the arguments of this paper appears to lie in the default nature of the conclusions under a pragmatic interpretation. It is argued here that loss should be taken to be logarithmic in the absence of knowledge of the true loss function. If one does know the true loss function L^* and this loss function turns out not to be logarithmic then one should arguably do something other than minimising worst-case expected logarithmic loss—one should minimise worst-case expected L^*-loss.

[10] L4 would need to be changed insofar as that it would need to be physical probability P^* rather than the agent's belief function B that determines whether sublanguages are independent. This change does not affect the formal results.

Under a non-pragmatic interpretation, on the other hand, one might argue that L1-4 characterises the correct measure of the inaccuracy of a belief in a true proposition, not a measure that is provisional in the sense that logarithmic loss is. Thus the conclusions of this paper are arguably firmer—less provisional—under a non-pragmatic construal.

§5.5. Questions for further research

We noted above that if one knows the true loss function L^* then one should arguably minimise worst-case expected L^*-loss. Grünwald and Dawid (2004) generalise standard entropy in a different direction to that pursued in this paper, in order to argue that minimising worst-case expected L^*-loss requires maximising entropy in their generalised sense. One interesting question for further research is whether one can generalise the notion of g-entropy in an analogous way, to try to show that minimising worst-case g-expected L^*-loss requires maximising g-entropy in this further generalised sense.

A second question concerns whether one can extend the discussion of belief over sentences in §3 to predicate, rather than propositional, languages. A third question is whether other justifications of logarithmic score can be used to justify logarithmic g-score—for example, is logarithmic g-score the only *local* strictly proper g-score? Fourth, we suspect that Theorem 25 can be further generalised. Finally, it would be interesting to investigate language invariance in more detail in order to test the conjecture at the end of §4.

Acknowledgements

This research was conducted as a part of the project *From objective Bayesian epistemology to inductive logic*. We are grateful to the UK Arts and Humanities Research Council for funding this research, and to Teddy Groves, Jeff Paris and three anonymous referees for very helpful comments.

A
Entropy of belief functions

Axiomatic characterizations of *standard* entropy on *probability functions* have featured heavily in the literature—see Csiszàr (2008). In this appendix we provide two characterizations of g-entropy on *belief functions* which closely resemble the original axiomatisation provided by Shannon (1948, §6). (We appeal to these characterisations in the proof of Proposition 55 in B.2.)

We shall need some new notation. Let $k \in \mathbb{N}$ and $x \in \mathbb{R}$, then denote by $x@k$ the tuple $\langle x, x, \ldots, x\rangle \in \mathbb{R}^k$. For $x \in \mathbb{R}$ and $\vec{y} \in \mathbb{R}^l$ we denote by $x \cdot \vec{y}$ the vector $\langle x \cdot y_1, \ldots, x \cdot y_l\rangle \in \mathbb{R}^l$. For a vector $\vec{x} \in \mathbb{R}^k$ let $|\vec{x}|_1 = x_1 + \ldots + x_k$. Assume in the following that all x_i and all y_{ij} are in $[0,1]$. Also, let k, l henceforth denote the number of components in $\vec{x}$ respectively $\vec{y}$.

Proposition 50 (*First characterisation*). *Let* $H(B) = \sum_{\pi\in\Pi} g(\pi) f(\pi, B)$ *where* $f(\pi, B) := h(B(F_1), \ldots, B(F_k))$ *for* $\pi = \{F_1, \ldots, F_k\}$ *and*

$$h : \bigcup_{k\geq 1} \{\langle x_1, \ldots, x_k\rangle \,:\, x_i \geq 0 \,\&\, \sum_{i=1}^{k} x_i \leq 1\} \longrightarrow [0, \infty).$$

Suppose also that the following conditions hold:

H1: *h is continuous;*

H2: *if $1 \leq t_1 < t_2 \in \mathbb{N}$ then $h(\frac{1}{t_1}@t_1) < h(\frac{1}{t_2}@t_2)$;*

H3: *if $0 < |\vec{x}|_1 \leq 1$ and if $|\vec{y}_i|_1 = 1$ for $1 \leq i \leq k$, then*

$$h(x_1 \cdot \vec{y}_1, \ldots, x_k \cdot \vec{y}_k) = h(x_1, \ldots, x_k) + \sum_{i=1}^{k} x_i h(\vec{y}_i);$$

H4: *$qh(\frac{1}{t}) = h(\frac{1}{t}@q)$ for $1 \leq q \leq t \in \mathbb{N}$;*

then $H(B) = -\sum_{\pi \in \Pi} g(\pi) \sum_{F \in \pi} B(F) \log B(F)$.

Proof: We first apply the proof of Paris (1994, pp. 77–78), which implies (using only H1, H2 and H3) that

$$h(\vec{x}) = -c \sum_{i=1}^{k} x_i \log x_i \tag{28}$$

for all $\vec{x}$ with $|\vec{x}|_1 = 1$, where $c \in \mathbb{R}_{>0}$ is some constant.

Now suppose $0 < |\vec{x}|_1 < 1$. Then with $y_i := \frac{x_i}{|\vec{x}|_1}$ we have $\vec{x} = |\vec{x}|_1 \cdot \vec{y}$ and $|\vec{y}|_1 = 1$. Thus

$$h(\vec{x}) = h(|\vec{x}|_1 \cdot \vec{y}) \overset{H3}{=} h(|\vec{x}|_1) + |\vec{x}|_1 h(\vec{y}) \overset{(28)}{=} h(|\vec{x}|_1) - |\vec{x}|_1 c \sum_{i=1}^{l} y_i \log y_i.$$

We will next show that $h(x) = -cx \log x$ for $x \in [0,1)$. Thus, note that $h(\frac{1}{t}) \overset{H4}{=} \frac{1}{t} h(\frac{1}{t}@t) \overset{(28)}{=} \frac{1}{t}(-ct\frac{1}{t}\log\frac{1}{t}) = -c\frac{1}{t}\log\frac{1}{t}$. For $1 \leq q \leq t \in \mathbb{N}$ we now find

$$\begin{aligned} -c\frac{1}{t}\log\frac{1}{t} = h(\frac{1}{t}) \overset{H4}{=} \frac{1}{q} h(\frac{1}{t}@q) = \frac{1}{q} h(\frac{q}{t} \cdot \frac{1}{q}@q) &\overset{H3}{=} \frac{1}{q}\Big(h(\frac{q}{t}) + \frac{q}{t} h(\frac{1}{q}@q)\Big) \\ &\overset{(28)}{=} \frac{1}{q}\Big(h(\frac{q}{t}) + \frac{q}{t}(-cq\frac{1}{q}\log\frac{1}{q})\Big). \end{aligned}$$

Thus

$$\begin{aligned} h(\frac{q}{t}) &= -c\frac{q}{t}\Big(\log(\frac{1}{t}) - \log(\frac{1}{q})\Big) \\ &= -c\frac{q}{t}\log\frac{q}{t}. \end{aligned}$$

Hence, h is of the claimed form for rational numbers in $(0,1]$. The continuity axiom H1 now guarantees that $h(x) = -cx\log x$ for all $x \in [0,1] \subset \mathbb{R}$. Putting our results together we obtain

$$\begin{aligned} h(\vec{x}) &= -c|\vec{x}|_1 \log|\vec{x}|_1 - c|\vec{x}|_1 \sum_{i=1}^{l} y_i \log y_i = -c|\vec{x}|_1(\sum_{i=1}^{l} y_i \log|\vec{x}|_1 + \sum_{i=1}^{l} y_i \log y_i) \\ &= -c|\vec{x}|_1 \sum_{i=1}^{l} y_i \log(|\vec{x}|_1 \cdot y_i) \\ &= -c\sum_{i=1}^{l} x_i \log x_i. \end{aligned}$$

Finally, note that h does satisfy all the axioms. The constant c can then be absorbed into the weighting function g to give $H(B) = -\sum_{\pi\in\Pi} g(\pi)\sum_{F\in\pi} B(F)\log B(F)$, as required. □

A tighter analysis reveals that the axiomatic characterization above may be weakened. We may replace H3 by the following two instances of H3:
A: If $|\vec{x}|_1 = 1$ and if $|\vec{y}_i|_1 = 1$ for $1 \le i \le k$, then

$$h(x_1 \cdot \vec{y}_1, \ldots, x_k \cdot \vec{y}_k) = h(x_1, \ldots, x_k) + \sum_{i=1}^{k} x_i h(\vec{y}_i).$$

B: If $0 < x < 1$ and if $|\vec{y}|_1 = 1$, then

$$h(x \cdot \vec{y}) = h(x) + x h(\vec{y}).$$

Property A is of course Shannon's original axiom H3. The axiom H3 used above is the straightforward generalization of Shannon's H3 to vectors $\vec{x}$ summing to less than one.

Proposition 51 (*Second characterisation*). *Let* $H(B) = \sum_{\pi\in\Pi} g(\pi) f(\pi,B)$ *where* $f(\pi,B) := h(B(F_1),...,B(F_k))$ *for* $\pi = \{F_1,...,F_k\}$ *and*

$$h : \bigcup_{k\ge1} \{\langle x_1, \ldots, x_k\rangle \ : \ x_i \ge 0 \ \& \ \sum_{i=1}^{k} x_i \le 1\} \longrightarrow [0,\infty).$$

Suppose also that the following conditions hold:

H1: h *is continuous;*

H2: *if* $1 \le t_1 < t_2 \in \mathbb{N}$ *then* $h(\frac{1}{t_1}@t_1) < h(\frac{1}{t_2}@t_2)$*;*

A: *if* $|\vec{x}|_1 = 1$ *and if* $|\vec{y}_i|_1 = 1$ *for* $1 \le i \le k$, *then*

$$h(x_1 \cdot \vec{y}_1, \ldots, x_k \cdot \vec{y}_k) = h(x_1, \ldots, x_k) + \sum_{i=1}^{k} x_i h(\vec{y}_i).$$

B : *if* $0 < x < 1$ *and if* $|\vec{y}|_1 = 1$, *then*

$$h(x \cdot \vec{y}) = h(x) + x h(\vec{y});$$

C: *for* $0 < x, y < 1$, *it holds that* $h(x \cdot y) = x h(y) + y h(x)$*;*

D: *for* $0 < x < 1$, *it holds that* $h(x) = h(x, 1-x) - h(1-x)$*;*

then $H(B) = -\sum_{\pi\in\Pi} g(\pi)\sum_{F\in\pi} B(F)\log B(F)$.

Proof: We shall again invoke the proof in Paris (1994, pp. 77–78) to show (using only H1, H2 and A) that

$$h(\vec{x}) = -c \sum_{i=1}^{k} x_i \log x_i \tag{29}$$

for all $\vec{x}$ with $|\vec{x}|_1 = 1$ and some constant $c \in \mathbb{R}_{>0}$.

Now suppose $0 < |\vec{x}|_1 < 1$. Then with $y_i := \frac{x_i}{|\vec{x}|_1}$ we have $\vec{x} = |\vec{x}|_1 \cdot \vec{y}$ and $|\vec{y}|_1 = 1$. Thus

$$h(\vec{x}) = h(|\vec{x}|_1 \cdot \vec{y}) \overset{H3}{=} h(|\vec{x}|_1) + |\vec{x}|_1 h(\vec{y}) \overset{(29)}{=} h(|\vec{x}|_1) - |\vec{x}|_1 c \sum_{i=1}^{l} y_i \log y_i.$$

As we have seen in the previous proof, it now only remains to show that $h(x) = -cx\log x$ for $x \in [0,1] \subset \mathbb{R}$.

We next show by induction that for all non-zero $t \in \mathbb{N}$, $h(\frac{1}{2^t}) = -c\frac{1}{2^t}\log\frac{1}{2^t}$.

The *base case* is immediate, observe that

$$h(\frac{1}{2}) \overset{D}{=} \frac{1}{2}h(\frac{1}{2},\frac{1}{2}) \overset{(29)}{=} -c\frac{1}{2}\log\frac{1}{2}.$$

Using the induction hypothesis (IH), the *inductive step* is straightforward too:

$$\begin{aligned} h(\frac{1}{2^t}) &\overset{C}{=} \frac{1}{2^{t-1}}h(\frac{1}{2}) + \frac{1}{2}h(\frac{1}{2^{t-1}}) \\ &\overset{IH}{=} -c\Big(\frac{1}{2^t}\log(\frac{1}{2}) + \frac{1}{2^t}\log(\frac{1}{2^{t-1}})\Big) \\ &= -c\frac{1}{2^t}\log\frac{1}{2^t}. \end{aligned}$$

We next show by induction on $t \geq 1$ that for all non-zero natural numbers $m < 2^t$, $h(\frac{m}{2^t}) = -c\frac{m}{2^t}\log\frac{m}{2^t}$.

For the *base case* simply note that $t = m = 1$ and thus

$$h(\frac{1}{2^1}) = -c\frac{1}{2}\log\frac{1}{2}.$$

The *inductive step* follows for $m < 2^{t-1}$:

$$\begin{aligned} h(\frac{m}{2^t}) &\overset{C}{=} \frac{m}{2^{t-1}}h(\frac{1}{2}) + \frac{1}{2}h(\frac{m}{2^{t-1}}) \\ &\overset{IH}{=} -c\frac{m}{2^{t-1}}\frac{1}{2}\log(\frac{1}{2}) - c\frac{1}{2}\frac{m}{2^{t-1}}\log(\frac{m}{2^{t-1}}) \\ &= -c\frac{m}{2^t}\log\frac{m}{2^t}. \end{aligned}$$

For $2^{t-1} < m < 2^t$ we find

$$\begin{aligned} h(\frac{m}{2^t}) &\overset{D}{=} h(\frac{m}{2^t}, \frac{2^t - m}{2^t}) - h(\frac{2^t - m}{2^t}) \\ &\overset{(29)}{=} -c\Big(\frac{m}{2^t}\log(\frac{m}{2^t}) + \frac{2^t - m}{2^t}\log(\frac{2^t - m}{2^t})\Big) - h(\frac{2^t - m}{2^t}) \\ &\overset{IH}{=} -c\Big(\frac{m}{2^t}\log(\frac{m}{2^t}) + \frac{2^t - m}{2^t}\log(\frac{2^t - m}{2^t})\Big) + c\frac{2^t - m}{2^t}\log(\frac{2^t - m}{2^t}) \\ &= -c\frac{m}{2^t}\log\frac{m}{2^t}. \end{aligned}$$

Since rational numbers of the form $\frac{m}{2^t}$ are dense in $[0,1] \subset \mathbb{R}$ we can use the continuity axiom H1 to conclude that h has to be of the desired form.

Finally, note that h does satisfy all the axioms. The constant c can then be absorbed into the weighting function g to give the required form of $H(B)$. □

We can combine B and C to form one single axiom H5 which implies B and C: *H5*: if $0 < x < 1$ and if $|\vec{y}|_1 \le 1$, then

$$h(x \cdot \vec{y}) = |\vec{y}|_1 h(x) + x h(\vec{y}).$$

Clearly, H5 is a natural way to generalize A to belief functions. It now follows easily that H1, H2, A, H5 and D jointly constrain h to $h(\vec{x}) = -c\sum_{i=1}^{k} x_i \log x_i$.

Although it is certainly possible to consider the g-entropy of a belief function, maximising standard entropy over $\mathbb{B}$—as opposed to $\mathbb{E} \subseteq \mathbb{P}$—has bizarre consequences. For $|\Omega| = 2$ we have that $\{B_z \in \mathbb{B} : z \in [0,1],\ B_z(\Omega) = z,\ B_z(\emptyset) = 1 - z,\ B_z(\omega_1) = B_z(\omega_2) = \frac{1}{e}\}$ is the set of entropy maximizers. This follows from considering the following optimization problem:

$$\begin{aligned} \text{maximize} \quad & -B(\omega_1)\log B(\omega_1) - B(\omega_2)\log B(\omega_2) \\ \text{subject to} \quad & 0 \le B(\emptyset), B(\Omega), B(\omega_1), B(\omega_2) \\ & B(\omega_1) + B(\omega_2) \le 1 \\ & B(\emptyset) + B(\Omega) \le 1 \\ & B(\emptyset) + B(\Omega) = 1 \text{ or } B(\omega_1) + B(\omega_2) = 1. \end{aligned}$$

Putting $B(\emptyset)+B(\Omega) = 1$ ensures that the last two constraints are satisfied and permits the choice of $B(\omega_1)$, $B(\omega_2)$ such that $B(\omega_1) + B(\omega_2) < 1$. For non-negative $B(\omega)$ we have that $-B(\omega)\log B(\omega)$ obtains the unique maximum at $B(\omega) = \frac{1}{e}$. The claimed optimality result follows.

It is worth pointing out that this phenomenon does not depend on the base of the logarithm. For $|\Omega| \ge 3$, however, intuition honed by considering entropy of probability functions does not lead one astray. For $|\Omega| \ge 3$, any belief function B with $B(\omega) = \frac{1}{|\Omega|}$ for $\omega \in \Omega$ does maximize standard entropy.

Similarly bizarre consequences also obtain in the case of other g-entropies. For $|\Omega| = 2$ and $g(\{\Omega\}) + g(\{\Omega, \emptyset\}) \ll g(\{\omega_1\}, \{\omega_2\})$, belief functions maximizing g-entropy satisfy $B(\omega_1) = B(\omega_2) = \frac{1}{e}$. To see this, simply note that for such g the optimum obtains for $B(\Omega) + B(\emptyset) = 1$.

For the proposition entropy for $|\Omega| = 2$, there are two entropy maximizers in $\mathbb{B}$. They are $B_1^\dagger(\emptyset) = B_1^\dagger(\Omega) = \frac{1}{2}$, $B_1^\dagger(\omega_1) = B_1^\dagger(\omega_2) = \frac{1}{e}$ and $B_2^\dagger(\emptyset) = B_2^\dagger(\Omega) = \frac{1}{e}$, $B_2^\dagger(\omega_1) = B_2^\dagger(\omega_2) = \frac{1}{2}$.

Thus, an agent adopting a belief function maximizing g-entropy over $\mathbb{B}$ may violate the probability norm. Furthermore, the agent may have to choose a belief function from finitely or infinitely many such non-probabilistic functions. For an agent minimizing worst-case g-expected loss these bizarre situations do not arise. From Theorem 24 and we know that for inclusive g, minimizing worst-case g-expected loss forces the agent to adopt a probability function which maximizes g-entropy over the set $\mathbb{E}$ of calibrated ***probability*** functions. By Corollary 10 this probability function is unique.

B
Properties of g-entropy maximisation

General properties of standard entropy (defined on probability functions) have been widely studied in the literature. Here we examine general properties of the g-entropy of a probability function, for $g \in \mathscr{G}$. We have already seen one difference between standard and g-entropy in Section §4: standard entropy satisfies language invariance; g-entropy in general need not. Surprisingly, language invariance seems to be an exception. Standard entropy and g-entropy behave in many respects in the same way.

B.1. Preserving the equivocator

For example, as we shall see now, if g is inclusive and symmetric then the probability function that is deemed most equivocal—i.e., the function, out of all probability functions, with maximum g-entropy—is the equivocator function $P_=$, which gives each state the same probability.

Definition 52 (*Equivocator-Preserving*). A weighting function g is called *equivocator-preserving*, if and only if $\arg\sup_{P\in\mathbb{P}} H_g(P) = P_=$.

That symmetry and inclusiveness are sufficient for g to be equivocator-preserving will follow from the following lemma:

Lemma 53. *For inclusive g, g is equivocator-preserving if and only if*

$$z(\omega) := \sum_{\substack{F\subseteq\Omega \\ \omega\in F}} \sum_{\substack{\pi\in\Pi \\ F\in\pi}} -g(\pi)(1-\log|\Omega|+\log|F|) = c,$$

for some constant c.

Proof: Recall from Proposition 8 that g-entropy is strictly concave on $\mathbb{P}$. Thus, every critical point in the interior of $\mathbb{P}$ is the unique maximiser of $H_g(\cdot)$ on $\mathbb{P}$.

Now consider the Lagrange function Lag:

$$\begin{aligned} Lag(P) &= \lambda(-1+\sum_{\omega\in\Omega} P(\omega)) + H_g(P) \\ &= \lambda(-1+\sum_{\omega\in\Omega} P(\omega)) + \sum_{\pi\in\Pi} -g(\pi) \sum_{F\in\pi} \Big(\sum_{\omega\in F} P(\omega)\Big)\Big(\log \sum_{\omega\in F} P(\omega)\Big). \end{aligned}$$

For fixed $\omega\in\Omega$ and $\pi\in\Pi$, denote by $F_{\omega,\pi}$ the unique $F\subseteq\Omega$ such that $\omega\in F$ and $F\in\pi$. Taking derivatives we obtain:

$$\frac{\partial}{\partial P(\omega)} Lag(P) = \lambda + \sum_{\pi\in\Pi} -g(\pi)(1+\log \sum_{\nu\in F_{\omega,\pi}} P(\nu)) \text{ for all } \omega\in\Omega.$$

Now, if $P_=$ maximises g-entropy, then for all $\omega\in\Omega$ the following must vanish:

$$\begin{aligned} \frac{\partial}{\partial P(\omega)} Lag(P_=) &= \lambda + \sum_{\pi\in\Pi} -g(\pi)(1+\log P_=(F_{\omega,\pi})) \\ &= \lambda + \sum_{\pi\in\Pi} -g(\pi)(1+\log\frac{|F_{\omega,\pi}|}{|\Omega|}) \\ &= \lambda + \sum_{\pi\in\Pi} -g(\pi)(1-\log|\Omega|+\log|F_{\omega,\pi}|) \\ &= \lambda + \sum_{\substack{F\subseteq\Omega \\ \omega\in F}} \sum_{\substack{\pi\in\Pi \\ F\in\pi}} -g(\pi)(1-\log|\Omega|+\log|F|). \end{aligned}$$

Since this expression has to vanish for all $\omega\in\Omega$, it does not depend on ω.

On the other hand, if g is such that

$$\sum_{\substack{F\subseteq\Omega\\ \omega\in F}}\sum_{\substack{\pi\in\Pi\\ F\in\pi}} -g(\pi)(1-\log|\Omega|+\log|F|)$$

does not depend on ω, then $P_=$ is a critical point of $Lag(P)$ and thus the entropy maximiser. □

Corollary 54. *If g is symmetric and inclusive then it is equivocator-preserving.*

Proof: By Lemma 53 we only need to show that

$$\sum_{\substack{F\subseteq\Omega\\ \omega\in F}}\sum_{\substack{\pi\in\Pi\\ F\in\pi}} -g(\pi)(1-\log|\Omega|+\log|F|)$$

does not depend on ω.

Denote by π_{ij} respectively F_{ij} the result of replacing ω_i by ω_j and vice versa in $\pi\in\Pi$, respectively $F\subseteq\Omega$. By the symmetry of g we have $g(\pi)=g(\pi_{ij})$. Since $|F|=|F_{ij}|$ we then find for all $\omega_i,\omega_j\in\Omega$,

$$\begin{aligned}\sum_{\substack{F\subseteq\Omega\\ \omega_i\in F}}\sum_{\substack{\pi\in\Pi\\ F\in\pi}} -g(\pi)(1-\log|\Omega|+\log|F|) &= \sum_{\substack{F\subseteq\Omega\\ \omega_i\in F}}\sum_{\substack{\pi\in\Pi\\ F\in\pi}} -g(\pi_{ij})(1-\log|\Omega|+\log|F_{ij}|)\\ &= \sum_{\substack{F\subseteq\Omega\\ \omega_i\in F}}\sum_{\substack{\pi\in\Pi\\ F_{ij}\in\pi}} -g(\pi)(1-\log|\Omega|+\log|F_{ij}|)\\ &= \sum_{\substack{F\subseteq\Omega\\ \omega_j\in F}}\sum_{\substack{\pi\in\Pi\\ F\in\pi}} -g(\pi)(1-\log|\Omega|+\log|F|).\end{aligned}$$

□

Are there are any non-symmetric, inclusive g that are equivocator preserving? We pose this as an interesting question for further research.

B.2. Updating

Next we show that there is widespread agreement between updating by conditionalisation and updating by g-entropy maximisation, a result to which we alluded in §5.

Proposition 55. *Suppose that $\mathbb{E}$ is the set of probability functions calibrated with evidence E. Let g be inclusive and $G\subseteq\Omega$ such that $\mathbb{E}'=\{P\in\mathbb{E} : P(G)=1\}\neq\emptyset$, where $\mathbb{E}'$ is the set of probability functions calibrated with evidence $E\cup\{G\}$. Then the following are equivalent:*

- $P^\dagger_E(\cdot|G)\in[\mathbb{E}]$
- $P^\dagger_{E\cup\{G\}}(\cdot)=P^\dagger_E(\cdot|G)$,

where $P^\dagger_E, P^\dagger_{E\cup\{G\}}$ are functions in $\mathbb{E},\mathbb{E}'$ respectively that maximise g-entropy.

Proof: First suppose that $P^\dagger_E(\cdot|G) \in [\mathbb{E}]$.

Observe that if $\mathbb{E}' = \mathbb{E}$, then there is nothing to prove. Thus suppose that $\mathbb{E}' \subset \mathbb{E}$. Hence, there exists a function $P \in \mathbb{E}$ with $P(\bar{G}) > 0$. By Proposition 70 inclusive g are open-minded, hence $P^\dagger_E(\bar{G}) > 0$.[11] So, $P^\dagger_E(\cdot|\bar{G})$ is well-defined.

Now let $P^\dagger_1 := P^\dagger_{E\cup\{G\}}$ and $P^\dagger := P^\dagger_E$. Then assume for contradiction that $P^\dagger_1(F) \neq P^\dagger(F|G)$ for some $F \subseteq \Omega$. By Corollary 10 the g-entropy maximiser $P^\dagger_1$ in $[\mathbb{E}']$ is unique, furthermore $P^\dagger(\cdot|G) \in [\mathbb{E}']$. It follows that:

$$\begin{aligned}\sum_{\pi\in\Pi} -g(\pi) \sum_{F'\in\pi} P^\dagger_1(F')\log P^\dagger_1(F') = H_g(P^\dagger_1)&\\ &> H_g(P^\dagger(\cdot|G))\\ &= \sum_{\pi\in\Pi} -g(\pi) \sum_{F'\in\pi} P^\dagger(F'|G)\log P^\dagger(F'|G).\end{aligned}$$

Now define $P'(\cdot) = P^\dagger(G)P^\dagger_1(\cdot|G) + P^\dagger(\bar{G})P^\dagger(\cdot|\bar{G})$. Since $[\mathbb{E}]$ is convex, $P^\dagger_1, P^\dagger(\cdot|G) \in [\mathbb{E}]$ and since $P^\dagger_1(\cdot|G) = P^\dagger_1$ we have that $P' \in [\mathbb{E}]$.

Using the above inequality we observe, using axiom A of Appendix A with $\vec{x} = \langle P^\dagger(G), P^\dagger(\bar{G})\rangle$, $\vec{y}_1 = \langle P^\dagger_1(F'|G) : F' \in \pi\rangle$ and $\vec{y}_2 = \langle P^\dagger(F'|G) : F' \in \pi\rangle$ that

$$\begin{aligned}H_g(P') &= \sum_{\pi\in\Pi} -g(\pi) \sum_{F'\in\pi} P'(F')\log P'(F')\\
&= \sum_{\pi\in\Pi} -g(\pi) \sum_{F'\in\pi} (P^\dagger(G)P^\dagger_1(F'|G) + P^\dagger(\bar{G})P^\dagger(F'|\bar{G}))\log(P^\dagger(G)P^\dagger_1(F'|G) + P^\dagger(\bar{G})P^\dagger(F'|\bar{G}))\\
&\overset{A}{=} \sum_{\pi\in\Pi} -g(\pi)\Big(P^\dagger(G)\log P^\dagger(G) + P^\dagger(\bar{G})\log P^\dagger(\bar{G})\Big)\\
&\qquad + \sum_{\pi\in\Pi} -g(\pi)\Big(P^\dagger(G)\sum_{F'\in\pi} P^\dagger_1(F'|G)\log P^\dagger_1(F'|G) + P^\dagger(\bar{G})\sum_{F'\in\pi} P^\dagger(F'|\bar{G})\log P^\dagger(F'|\bar{G})\Big)\\
&= \sum_{\pi\in\Pi} -g(\pi)\Big(P^\dagger(G)\log P^\dagger(G) + P^\dagger(\bar{G})\log P^\dagger(\bar{G})\Big)\\
&\qquad + \sum_{\pi\in\Pi} -g(\pi)\Big(P^\dagger(G)\sum_{F'\in\pi} P^\dagger_1(F')\log P^\dagger_1(F') + P^\dagger(\bar{G})\sum_{F'\in\pi} P^\dagger(F'|\bar{G})\log P^\dagger(F'|\bar{G})\Big)\\
&> \sum_{\pi\in\Pi} -g(\pi)\Big(P^\dagger(G)\log P^\dagger(G) + P^\dagger(\bar{G})\log P^\dagger(\bar{G})\Big)\\
&\qquad + \sum_{\pi\in\Pi} -g(\pi)\Big(P^\dagger(G)\sum_{F'\in\pi} P^\dagger(F'|G)\log P^\dagger(F'|G) + P^\dagger(\bar{G})\sum_{F'\in\pi} P^\dagger(F'|\bar{G})\log P^\dagger(F'|\bar{G})\Big)\\
&= H_g(P^\dagger).\end{aligned}$$

Our above calculation contradicts that $P^\dagger$ maximises g-entropy over $[\mathbb{E}]$. Thus, $P^\dagger_1(\cdot) = P^\dagger(\cdot|G)$.

Conversely, suppose that $P^\dagger_E(\cdot|G) = P^\dagger_{E\cup\{G\}}(\cdot)$. Now simply observe $P^\dagger_E(\cdot|G) \in [\mathbb{E}'] \subseteq [\mathbb{E}]$. □

Theorem 49. *Suppose that convex and closed $\mathbb{E}$ is the set of probability functions calibrated with evidence E, and $\mathbb{E}'$ is the set of probability functions calibrated with evidence $E \cup \{G\}$. Also suppose that $P^\dagger_E, P^\dagger_{E\cup\{G\}}$ are functions in $\mathbb{E}, \mathbb{E}'$ respectively that maximise g-entropy for some fixed $g \in \mathcal{G}_{\text{inc}} \cup \{g_\Omega\}$. If*

[11] Note that the proof of Proposition 70 does not itself depend on Proposition 55.

(i) $G \subseteq \Omega$,

(ii) *the only constraints imposed by* $E \cup \{G\}$ *are the constraints imposed by* E *together with the constraint* $P(G) = 1$,

(iii) *the constraints in (ii) are consistent, and*

(iv) $P^{\dagger}_{E}(\cdot|G) \in \mathbb{E}$,

then $P^{\dagger}_{E\cup\{G\}}(F) = P^{\dagger}_{E}(F|G)$ *for all* $F \subseteq \Omega$.

Proof: For $g \in \mathscr{G}_{\text{inc}}$ this follows directly from Proposition 55. Simply note that $\mathbb{E} = [\mathbb{E}]$ and thus $P^{\dagger}_{E}(\cdot|G) \in [\mathbb{E}]$.

The proof of Proposition 55 also goes through for $g = g_{\Omega}$. This follows from the fact that all the ingredients in the proof—open-mindedness, uniqueness of the g-entropy maximiser on a convex set $\mathbb{E}$ and the axiomatic characterizations in Appendix A—also hold for standard entropy. □

This extends Seidenfeld's result for standard entropy, Theorem 48, to arbitrary convex sets $\mathbb{E} \subseteq \mathbb{P}$ and also to inclusive weighting functions.

B.3. Paris-Vencovská Properties

The following eight principles have played a central role in axiomatic characterizations of the maximum entropy principle by Paris and Vencovská—c.f., Paris and Vencovská (1990); Paris (1994); Paris and Vencovská (1997); Paris (1998). The first seven principles were first put forward in Paris and Vencovská (1990). Paris (1998) views all eight principles as following from the following single common-sense principle: "Essentially similar problems should have essentially similar solutions."

While Paris and Vencovská mainly considered linear constraints, we shall consider arbitrary convex sets $\mathbb{E}, \mathbb{E}_1$. Adopting their definitions and using our notation we investigate the following properties:

Definition 56 (*1: Equivalence*). $P^{\dagger}$ only depends on $\mathbb{E}$ and not on the constraints that give rise to $\mathbb{E}$.

This clearly holds for every weighting function g.

Definition 57 (*2: Renaming*). Let per be an element of the permutation group on $\{1,\ldots,|\Omega|\}$. For a proposition $F \subseteq \Omega$ with $F = \{\omega_{i_1},\ldots,\omega_{i_k}\}$ define $per(F) = \{\omega_{per(i_1)},\ldots,\omega_{per(i_k)}\}$. Next let $per(B(F)) = B(per(F))$ and $per(\mathbb{E}) = \{per(P) : P \in \mathbb{E}\}$. Then g satisfies renaming if and only if $P^{\dagger}_{\mathbb{E}}(F) = P^{\dagger}_{per(\mathbb{E})}(per(F))$.

Proposition 58. *If* g *is inclusive and symmetric then* g *satisfies renaming.*

Proof: For $\pi \in \Pi$ with $\pi = \{F_{i_1},\ldots,F_{i_f}\}$ define $per(\pi) = \{per(F_{i_1}),\ldots,per(F_{i_f})\}$.

Using that g is symmetric for the second equality we find

$$\begin{aligned}
H_g(P) &= -\sum_{\pi\in\Pi} g(\pi) \sum_{F\in\pi} P(F)\log P(F) \\
&= -\sum_{\pi\in\Pi} g(per^{-1}(\pi)) \sum_{F\in\pi} P(F)\log P(F) \\
&= -\sum_{\pi\in\Pi} g(\pi) \sum_{F\in per(\pi)} P(F)\log P(F) \\
&= -\sum_{\pi\in\Pi} g(\pi) \sum_{F\in\pi} P(per(F))\log P(per(F)) \\
&= H_g(per(P)).
\end{aligned}$$

Thus $P^\dagger_{per(\mathbb{E})} = per(P^\dagger)$ and hence $P^\dagger_{per(\mathbb{E})}(per(F)) = per(P^\dagger)(per(F)) = P^\dagger(F)$. □

Weighting functions g satisfying the renaming property satisfy a further symmetry condition, as we shall see now.

Definition 59 (*Symmetric complement*). For $P \in \mathbb{P}$ define the *symmetric complement* of P with respect to A_i, denoted by $\sigma_i(P)$, as follows:

$$\begin{aligned}
&\sigma_i(P)(\pm A_1 \wedge \ldots \wedge \pm A_{i-1} \wedge \pm A_i \wedge \pm A_{i+1} \wedge \ldots \wedge \pm A_n) \\
&\quad := P(\pm A_1 \wedge \ldots \wedge \pm A_{i-1} \wedge \mp A_i \wedge \pm A_{i+1} \wedge \ldots \wedge \pm A_n),
\end{aligned}$$

i.e., $\sigma_i(P)(\omega) = P(\omega')$ where ω' is ω but with A_i negated. A function $P \in \mathbb{P}$ is called *symmetric with respect to* A_i if and only if $P = \sigma_i(P)$.

We call $\mathbb{E} \subseteq \mathbb{P}$ symmetric with respect to A_i just when the following condition holds: $P \in [\mathbb{E}]$ if and only if $\sigma_i(P) \in [\mathbb{E}]$.

Corollary 60. *For all symmetric and inclusive g and all* $\mathbb{E}$ *that are symmetric with respect to* A_i *it holds that*

$$P^\dagger = \sigma_i(P^\dagger).$$

Thus, if $\mathbb{E}$ is symmetric with respect to A_i, so is $P^\dagger$.

Proof: Since g is symmetric and inclusive there is some function $\gamma : \mathbb{N} \to \mathbb{R}_{>0}$ such that $H_g(P) = \sum_{F\subseteq\Omega} -\gamma(|F|)P(F)\log P(F)$ for all $P \in \mathbb{P}$. Hence,

$$\begin{aligned}
H_g(P^\dagger) &= \sum_{F\subseteq\Omega} -\gamma(|F|)P^\dagger(F)\log P^\dagger(F) \\
&= \sum_{F\subseteq\Omega} -\gamma(|F|)\cdot\sigma_i(P^\dagger)(F)\cdot\log(\sigma_i(P^\dagger)(F)) \\
&= H_g(\sigma_i(P^\dagger)).
\end{aligned}$$

Since $\mathbb{E}$ is symmetric with respect to A_i we have that $\sigma_i(P^\dagger) \in [\mathbb{E}]$. So, if $P^\dagger \neq \sigma_i(P^\dagger)$, then there are two different probability functions in $[\mathbb{E}]$ which both have maximum entropy. This contradicts the uniqueness of the g-entropy maximiser (Corollary 10). □

This Corollary explains the symmetries exhibited in the tables in the proof of Proposition 42. Since in that proof $\mathbb{E}$ is symmetric with respect to A_3, the proposition entropy and the partition entropy maximisers are symmetric with respect to A_3. Thus, $P^\dagger_{\mathscr{P}\Omega,\mathscr{L}'}(\omega \wedge A_3) = P^\dagger_{\mathscr{P}\Omega,\mathscr{L}'}(\omega \wedge \neg A_3)$ and $P^\dagger_{\Pi,\mathscr{L}'}(\omega \wedge A_3) = P^\dagger_{\Pi,\mathscr{L}'}(\omega \wedge \neg A_3)$ for all $\omega \in \Omega$.

Definition 61 (*3: Irrelevance*). Let $\mathbb{P}_1, \mathbb{P}_2$ be the sets of probability functions on disjoint $\mathscr{L}_1, \mathscr{L}_2$ respectively. Then Irrelevance holds if, for $\mathbb{E}_1 \subseteq \mathbb{P}_1$ and $\mathbb{E}_2 \subseteq \mathbb{P}_2$, we have that $P^\dagger_{\mathbb{E}_1}(F \times \Omega_2) = P^\dagger_{\mathbb{E}_1 \times \mathbb{E}_2}(F \times \Omega_2)$ for all propositions F of $\mathscr{L}_1$, where $P^\dagger_{\mathbb{E}_1}, P^\dagger_{\mathbb{E}_1 \times \mathbb{E}_2}$ are the g-entropy maximisers on $\mathscr{L}_1 \cup \mathscr{L}_2$ with respect to $\mathbb{E}_1 \times \mathbb{P}_2$, respectively $\mathbb{E}_1 \times \mathbb{E}_2$.

Proposition 62. *Neither the partition nor the proposition weighting satisfy irrelevance.*

Proof: Let $\mathscr{L}_1 = \{A_1, A_2\}$, $\mathscr{L}_2 = \{A_3\}$, $\mathbb{E}_1 = \{P \in \mathbb{P}_1 : P(A_1 \wedge A_2) + 2P(\neg A_1 \wedge \neg A_2) = 0.2\}$ and $\mathbb{E}_2 = \{P \in \mathbb{P}_2 : P(A_3) = 0.1\}$. Then with $\omega_1 = \neg A_1 \wedge \neg A_2 \wedge \neg A_3$, $\omega_2 = \neg A_1 \wedge \neg A_2 \wedge A_3$ and so on we find:

	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7	ω_8
$P^\dagger_{\Pi,\mathbb{E}_1}$	0.0142	0.0142	0.2071	0.2071	0.2071	0.2071	0.0715	0.0715
$P^\dagger_{\Pi,\mathbb{E}_1 \times \mathbb{E}_2}$	0.0312	0.0004	0.3692	0.0466	0.3692	0.0466	0.1304	0.0064
$P^\dagger_{\mathscr{P}\Omega,\mathbb{E}_1}$	0.0050	0.0050	0.2025	0.2025	0.2025	0.2025	0.0901	0.0901
$P^\dagger_{\mathscr{P}\Omega,\mathbb{E}_1 \times \mathbb{E}_2}$	0.0211	$6.2 \cdot 10^{-9}$	0.3606	0.0500	0.3606	0.0500	0.1577	$2.3 \cdot 10^{-6}$

Now simply note that for instance

$$\begin{aligned} P^\dagger_{\Pi,\mathbb{E}_1}(\neg A_1 \wedge \neg A_2) &= P^\dagger_{\Pi,\mathbb{E}_1}(\omega_1) + P^\dagger_{\Pi,\mathbb{E}_1}(\omega_2) \\ &\neq P^\dagger_{\Pi,\mathbb{E}_1 \times \mathbb{E}_2}(\omega_1) + P^\dagger_{\Pi,\mathbb{E}_1 \times \mathbb{E}_2}(\omega_2) = P^\dagger_{\Pi,\mathbb{E}_1 \times \mathbb{E}_2}(\neg A_1 \wedge \neg A_2). \end{aligned}$$

(As we are going to see in Proposition 70 none of the values in the table can be zero. So the small numerical values found by computer approximation are not artifacts of the approximations involved.) □

Definition 63 (*4: Relativisation*). Let $\emptyset \subset F \subset \Omega$ and $\mathbb{E} = \{P \in \mathbb{P} : P(F) = z\} \cap \mathbb{E}_1 \cap \mathbb{E}_2$ and $\mathbb{E}' = \{P \in \mathbb{P} : P(F) = z\} \cap \mathbb{E}_1 \cap \mathbb{E}'_2$ where $\mathbb{E}_1$ is determined by a set of constraints on the $P(G)$ with $G \subseteq F$ and the $\mathbb{E}_2, \mathbb{E}'_2$ are determined by a set of constraints on the $P(G)$ with $G \subseteq \bar{F}$. Then $P^\dagger_{\mathbb{E}}(G) = P^\dagger_{\mathbb{E}'}(G)$ for all $G \subseteq F$.

Proposition 64. *Neither the partition not the proposition weighting satisfy relativisation.*

Proof: Let $|\Omega| = 8$, $F = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5\}$, $P(F) = 0.5$ and put $\mathbb{E}_1 = \{P \in \mathbb{P} : P(\omega_1) + 2P(\omega_2) + 3P(\omega_3) + 4P(\omega_4) = 0.2\}, \mathbb{E}_2 = \mathbb{P}, \mathbb{E}'_2 = \{P \in \mathbb{P} : P(\omega_6) + 2P(\omega_7) + 3P(\omega_8) = 0.7\}$. Then $P^\dagger_{\Pi,\mathbb{E}}$ and $P^\dagger_{\Pi,\mathbb{E}'}$ differ substantially on three out of five $\omega \in F$, as do $P^\dagger_{\mathscr{P}\Omega,\mathbb{E}}$ and $P^\dagger_{\mathscr{P}\Omega,\mathbb{E}'}$, as can be seen from the following table:

	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7	ω_8
$P^\dagger_{\Pi,\mathbb{E}}$	0.1251	0.0308	0.0041	0.0003	0.3398	0.1667	0.1667	0.1667
$P^\dagger_{\Pi,\mathbb{E}'}$	0.1242	0.0312	0.0041	0.0003	0.3402	0.3356	0.1288	0.0356
$P^\dagger_{\mathscr{P}\Omega,\mathbb{E}}$	0.1523	0.0239	$5.5 \cdot 10^{-7}$	$6.8 \cdot 10^{-9}$	0.3239	0.1667	0.1667	0.1667
$P^\dagger_{\mathscr{P}\Omega,\mathbb{E}'}$	0.1495	0.0252	$7.0 \cdot 10^{-7}$	$7.6 \cdot 10^{-9}$	0.3252	0.3252	0.1495	0.0252

□

Definition 65 (*5: Obstinacy*). If $\mathbb{E}_1$ is a subset of $\mathbb{E}$ such that $P^\dagger_{\mathbb{E}} \in [\mathbb{E}_1]$, then $P^\dagger_{\mathbb{E}} = P^\dagger_{\mathbb{E}_1}$.

Proposition 66. *If g is inclusive then it satisfies the obstinacy principle.*

Proof: This follows directly from the definition of $P^\dagger_{\mathbb{E}}$. □

Definition 67 (*6: Independence*). If $\mathbb{E} = \{P \in \mathbb{P} \mid P(A_1 \wedge A_3) = \alpha,\ P(A_2 \wedge A_3) = \beta,\ P(A_3) = \gamma\}$, then for $\gamma > 0$ it holds that $P^\dagger(A_1 \wedge A_2 \wedge A_3) = \frac{\alpha\beta}{\gamma}$.

Proposition 68. *Neither the partition entropy nor the proposition weighting satisfy independence.*

Proof: Let $\mathscr{L} = \{A_1, A_2, A_3\}$, $\alpha = 0.2$, $\beta = 0.35$, $\gamma = 0.6$, then

$$P^\dagger_\Pi(A_1 \wedge A_2 \wedge A_3) = 0.1197 \neq 0.1167 = \frac{0.2 \cdot 0.35}{0.6}$$

and

$$P^\dagger_{\mathscr{P}\Omega}(A_1 \wedge A_2 \wedge A_3) = 0.1237 \neq 0.1167 = \frac{0.2 \cdot 0.35}{0.6}.$$

□

Definition 69 (*7: Open-mindedness*). A weighting function g is *open-minded*, if and only if for all $\mathbb{E}$ and all $\emptyset \subseteq F \subseteq \Omega$ it holds that $P^\dagger(F) = 0$ if and only if $P(F) = 0$ for all $P \in \mathbb{E}$.

Proposition 70. *Any inclusive g is open-minded.*

Proof: First, observe that $P(F) = 0$ for all $P \in \mathbb{E}$, if and only if $P(F) = 0$ for all $P \in [\mathbb{E}]$.

Now note that if $P(F) = 0$ for all $P \in [\mathbb{E}]$, then $P^\dagger_g(F) = 0$, since $P^\dagger_g \in [\mathbb{E}]$. On the other hand, if there exists an $F \subseteq \Omega$ such that $P^\dagger_g(F) = 0 < P(F)$ for some $P \in [\mathbb{E}]$, then $S^{\log}_g(P, P^\dagger_g) = \infty > H_g(P^\dagger_g)$. Thus, adopting $P^\dagger_g$ exposes one to an infinite loss and by Theorem 24 adopting the g-entropy maximiser exposes one to the finite loss $H_g(P^\dagger_g)$. Contradiction. Thus, $P^\dagger_g(F) > 0$.

Overall, $P^\dagger_g(F) = 0$ if and only if $P(F) = 0$ for all $P \in [\mathbb{E}]$. □

Definition 71 (*8: Continuity*). Let us recall the definition of the *Blaschke metric* Δ between two convex sets $\mathbb{E}, \mathbb{E}_1 \subseteq \mathbb{P}$:

$$\begin{aligned}\Delta(\mathbb{E}, \mathbb{E}_1) = \inf\{\delta \mid\ & \forall P \in \mathbb{E}\, \exists P_1 \in \mathbb{E}_1 : |P, P_1| \le \delta \\ & \&\ \forall P_1 \in \mathbb{E}_1\, \exists P \in \mathbb{E} : |P, P_1| \le \delta\},\end{aligned}$$

where $|\cdot,\cdot|$ is the usual Euclidean metric between elements of $\mathbb{R}^{|\Omega|}$. g satisfies continuity if and only if the function $\arg\sup_{P \in \mathbb{E}} H_g(P)$ is continuous in the Blaschke metric.

Proposition 72. *Any inclusive g satisfies the continuity property.*

Proof: Since the g-entropy is strictly concave, see Proposition 8, we may apply Theorem 7.5 in (Paris, 1994, p. 91). Thus if $\mathbb{E}$ is determined by finitely many linear constraints then g satisfies continuity. Paris (1994) credits I. Maung for the proof of the theorem.

Now let $\mathbb{E} \subseteq \mathbb{P}$ be an arbitrary convex set. Note that we can approximate $\mathbb{E}$ arbitrarily closely by two sequences $\mathbb{E}_t, \mathbb{E}^t$ where each member of the sequences is determined by finitely many linear constraints such that $\mathbb{E}_t \subseteq \mathbb{E}_{t+1} \subseteq \mathbb{E} \subseteq \mathbb{E}^{t+1} \subseteq \mathbb{E}^t$. By this subset relation we have $\sup_{P \in \mathbb{E}_t} H_g(P) \leq \sup_{P \in \mathbb{E}} H_g(P) \leq \sup_{P \in \mathbb{E}^t} H_g(P)$. With $P^{\dagger_t} := \arg\sup_{P \in \mathbb{E}_t} H_g(P)$ and $P^{\dagger^t} := \arg\sup_{P \in \mathbb{E}^t} H_g(P)$ we have $\lim_{t \to \infty} P^{\dagger_t} = \lim_{t \to \infty} P^{\dagger^t}$ by Maung's theorem.

Since $\mathbb{E}^t$ converges to $\mathbb{E}_t$ in the Blaschke metric we have by Maung's theorem that $\lim_{t \to \infty} \sup_{P \in \mathbb{E}_t} H_g(P) = \lim_{t \to \infty} \sup_{P \in \mathbb{E}^t} H_g(P) = \sup_{P \in \mathbb{E}} H_g(P)$. Note that $\lim_{t \to \infty} P^{\dagger_t} \in [\mathbb{E}]$. Moreover, since $\mathbb{E}$ is convex, H_g is strictly concave and since $\mathbb{E}_t$ converges to $\mathbb{E}$ we have $\lim_{t \to \infty} H_g(P^{\dagger_t}) = \sup_{P \in \mathbb{E}} H_g(P)$. By the uniqueness of the g-entropy maximiser on $\mathbb{E}$ we thus find $\lim_{t \to \infty} P^{\dagger_t} = P^\dagger$, $\lim_{t \to \infty} P^{\dagger^t} = P^\dagger$ and $\lim_{t \to \infty} P^{\dagger_t} = \lim_{t \to \infty} P^{\dagger^t}$.

Since the sets determined by finitely many linear constraints are dense in the set of convex $\mathbb{E} \subseteq \mathbb{P}$ we can use a standard approximation argument yielding that $\arg\sup_{P \in \mathbb{E}} H_g(P)$ is continuous in the Blaschke metric on the set of convex $\mathbb{E} \subseteq \mathbb{P}$. □

B.4. The topology of g-entropy

We have so far investigated g-entropy for fixed $g \in \mathcal{G}$. We now briefly consider location and shape of the set of g-entropy maximisers.

For standard entropy maximisation and g-entropy maximisation with inclusive and symmetric g the respective maximisers all obtain at $P_=$, if $P_= \in [\mathbb{E}]$; cf Corollary 54.

If $P_= \notin [\mathbb{E}]$, then the maxima all obtain at the boundary of $\mathbb{E}$ "facing" $P_=$. To make this latter observation precise we denote for $P, P' \in \mathbb{P}$ the line segment in $\mathbb{P}$ which connects P with P', end points included, by $\overline{PP'}$.

Proposition 73 (*g-entropy is maximised at the boundary*). *For inclusive and symmetric g, $\overline{P_= P^\dagger} \cap [\mathbb{E}] = \{P^\dagger\}$.*

Proof: If $P_= \in [\mathbb{E}]$, then $P^\dagger = P_=$, by Corollary 54.

If $P_= \notin [\mathbb{E}]$, suppose that there exists a $P' \in \overline{P_= P^\dagger} \cap [\mathbb{E}]$ different from $P^\dagger$. Then by the concavity of g-entropy on $\mathbb{P}$ (Proposition 8) and the equivocator-preserving property (Corollary 54) we have $H_g(P_=) > H_g(P') > H_g(P^\dagger)$. By the convexity of $[\mathbb{E}]$ and Proposition 8 we have $H_g(P^\dagger) > H_g(P)$ for all $P \in [\mathbb{E}] \setminus \{P^\dagger\}$. Contradiction. □

We saw in Theorem 39 that for a particular sequence g_t converging to g_Ω, $P^\dagger_{g_t}$ converges to $P^\dagger_\Omega$. We shall now show that this is an instance of a more general phenomenon. We will demonstrate that $P^\dagger_g$ varies continuously for continuous changes in g for $g \in \mathcal{G}$.

Proposition 74 (*Continuity of g-entropy maximisation*). *For all $\mathbb{E}$, the function*

$$\arg\sup_{P \in \mathbb{E}} H_{(\cdot)}(P) : \mathcal{G} \longrightarrow [\mathbb{E}], \quad g \mapsto P^\dagger_g$$

is continuous on $\mathscr{G}$.

Proof: Consider a sequence $(g_t)_{t\in\mathbb{N}} \subseteq \mathscr{G}$ converging to some $g \in \mathscr{G}$. We need to show that $P^\dagger_{g_t}$ converges to $P^\dagger_g$.

From g_t converging to g it easily follows that $H_{g_t}(P)$ converges to $H_g(P)$ for all $P \in \mathbb{P}$.

Since g-entropy is strictly concave we have that for every $P' \in [\mathbb{E}] \setminus \{P^\dagger_g\}$ there exists some $\epsilon > 0$ such that $H_g(P') + \epsilon = H_g(P^\dagger_g)$. By the fact that $H_{g_t}(P)$ converges to $H_g(P)$ for all P we find that $H_{g_t}(P') + \frac{\epsilon}{2} < H_{g_t}(P^\dagger_g)$ for all t which are greater than some $T \in \mathbb{N}$.

Since $H_{g_t}(P^\dagger_g) \le H_{g_t}(P^\dagger_{g_t})$ it follows that P' cannot be a point of accumulation of the sequence $(P^\dagger_{g_t})_{t\in\mathbb{N}}$.

The sequence $P^\dagger_{g_t}$ takes values in the compact set $[\mathbb{E}]$, so it has at least one point of accumulation. We have demonstrated above that $P^\dagger_g$ is the only possible point of accumulation. Hence, $P^\dagger_g$ is the only point of accumulation and therefore the limit of this sequence. □

The continuity of g-entropy maximisation will be instrumental in proving the next proposition which asserts that the g-entropy maximisers are clustered together.

Proposition 75. *For any* $\mathbb{E}$, *if* $\mathscr{G} \subseteq \mathscr{G}_{\text{inc}}$ *is path-connected then the set* $\{P^\dagger_g : g \in \mathscr{G}\}$ *is path-connected.*

Proof: By Proposition 74 the map $\arg\sup_{P\in\mathbb{E}} H_{(\cdot)}(P)$ is continuous. The image of a path-connected set under a continuous map is path-connected. □

Corollary 76. *For all* $\mathbb{E}$, *the sets* $\{P^\dagger_g : g \in \mathscr{G}_{\text{inc}}\}$ *and* $\{P^\dagger_g : g \in \mathscr{G}_0\}$ *are path-connected.*

Proof: $\mathscr{G}_{\text{inc}}$ and $\mathscr{G}_0$ are convex, thus they are path-connected. Now apply Proposition 75. □

It is in general not the case that a convex combination of weighting functions generates a convex combination of the corresponding g-entropy maximisers:

Proposition 77. *For a convex combination of weighting functions* $g = \lambda g_1 + (1-\lambda) g_2$ *in general it fails to hold that* $P^\dagger_g = \lambda P^\dagger_{g_1} + (1-\lambda) P^\dagger_{g_2}$. *Moreover, in general* $P^\dagger_g \notin \overline{P^\dagger_{g_1} P^\dagger_{g_2}}$.

Proof: Let $g_1 = g_\Pi$, $g_2 = g_{\mathscr{P}\Omega}$ and $\lambda = 0.3$. Then for a language $\mathscr{L}$ with two propositional variables and $\mathbb{E} = \{P \in \mathbb{P} : P(\omega_1) + 2P(\omega_2) + 3P(\omega_3) + 4P(\omega_4) = 1.7\}$ we can see from the following table that$P^\dagger_{0.3g_\Pi + 0.7g_{\mathscr{P}\Omega}} \neq 0.3P^\dagger_\Pi + 0.7P^\dagger_{\mathscr{P}\Omega}$.

	ω_1	ω_2	ω_3	ω_4
$P^\dagger_\Pi$	0.5331	0.2841	0.1324	0.0504
$P^\dagger_{\mathscr{P}\Omega}$	0.5192	0.3008	0.1408	0.0392
$0.3P^\dagger_\Pi + 0.7P^\dagger_{\mathscr{P}\Omega}$	0.5234	0.2958	0.1383	0.0426
$P^\dagger_{0.3g_\Pi+0.7g_{\mathscr{P}\Omega}}$	0.5272	0.2915	0.1353	0.0459
$\frac{P^\dagger_{\mathscr{P}\Omega} - P^\dagger_{0.3g_\Pi+0.7g_{\mathscr{P}\Omega}}}{P^\dagger_{\mathscr{P}\Omega} - P^\dagger_\Pi}$	0.5755	0.5569	0.6429	0.6036

If $P^\dagger_{0.3g_\Pi+0.7g_{\mathscr{P}\Omega}}$ were in $\overline{P^\dagger_\Pi P^\dagger_{\mathscr{P}\Omega}}$, then the last line of the above table would be constant for all $\omega \in \Omega$. As we can see, the values in the last line do vary. □

C
Level of generalisation

In this section we shall show that the generalisation of entropy and score used in the text above is essentially the right one. We shall do this by defining broader notions of entropy and score of which the g-entropy and g-score are special cases, and showing that entropy maximisation only coincides with minimisation of worst-case score in the special case of g-entropy and g-score as they are defined above.

We will focus on the case of belief over propositions; belief over sentences behaves similarly. Our broader notions will be defined relative to a weighting $\gamma : \mathscr{P}\Omega \longrightarrow \mathbb{R}_{\geq 0}$ of propositions rather than a weighting $g : \Pi \longrightarrow \mathbb{R}_{\geq 0}$ of partitions.

Definition 78 (*γ-entropy*). Given a function $\gamma : \mathscr{P}\Omega \longrightarrow \mathbb{R}_{\geq 0}$, the γ-entropy of a normalised belief function is defined as

$$H_\gamma(B) := -\sum_{F\subseteq\Omega} \gamma(F)B(F)\log B(F).$$

Definition 79 (*γ-score*). Given a loss function L and a function $\gamma : \mathscr{P}\Omega \longrightarrow \mathbb{R}_{\geq 0}$, the γ-expected loss function or γ-scoring rule or simply γ-score is $S^L_\gamma : \mathbb{P} \times \langle\mathbb{B}\rangle \longrightarrow [-\infty,\infty]$ such that $S^L_\gamma(P,B) = \sum_{F\subseteq\Omega} \gamma(F)P(F)L(F,B)$.

Definition 80 (*Equivalent to a weighting of partitions*). A weighting of propositions $\gamma : \mathscr{P}\Omega \longrightarrow \mathbb{R}_{\geq 0}$ is *equivalent to a weighting of partitions* if there exists a function $g : \Pi \longrightarrow \mathbb{R}_{\geq 0}$ such that for all $F \subseteq \Omega$,

$$\gamma(F) = \sum_{\substack{\pi\in\Pi \\ F\in\pi}} g(\pi).$$

We see then that the notions of g-entropy and g-score coincide with those of γ-entropy and γ-score just when the weightings of propositions γ are equivalent to weightings of partitions. Next we extend the notion of inclusivity to our more general weighting functions:

Definition 81 (*Inclusive weighting of propositions*). A weighting of propositions $\gamma : \mathscr{P}\Omega \longrightarrow \mathbb{R}_{\geq 0}$ is *inclusive* if $\gamma(F) > 0$ for all $F \subseteq \Omega$.

We shall also consider a slight generalisation of strict propriety (cf., footnote 6):

Definition 82 (*Strictly $\mathbb{X}$-proper γ-score*). For $P \subseteq \mathbb{X} \subseteq \langle\mathbb{B}\rangle$, a γ-score $S^L_\gamma : \mathbb{P} \times \langle\mathbb{B}\rangle \longrightarrow [-\infty,\infty]$ is *strictly $\mathbb{X}$-proper* if for all $P \in \mathbb{P}$, the restricted function $S^L_\gamma(P,\cdot) : \mathbb{X} \longrightarrow [-\infty,\infty]$ has a unique global minimum at $B = P$. A γ-score is *strictly proper* if it is strictly $\langle\mathbb{B}\rangle$-proper. A γ-score is *merely $\mathbb{X}$-proper* if for some P this minimum at $B = P$ is not the only minimum.

Note that if a γ-score is strictly $\mathbb{X}$-proper then it is strictly $\mathbb{Y}$-proper for $\mathbb{P} \subseteq \mathbb{Y} \subseteq \mathbb{X}$. Thus if it is strictly proper it is also strictly $\mathbb{B}$-proper and strictly $\mathbb{P}$-proper.

Proposition 83. *Logarithmic γ-score $S^{\log}_\gamma(P,B)$ is non-negative and convex as a function of $B \in \langle\mathbb{B}\rangle$. For inclusive γ, convexity is strict, i.e., $S^{\log}_\gamma(P,\lambda B_1 + (1-\lambda)B_2) < \lambda S^{\log}_\gamma(P,B_1) + (1-\lambda)S^{\log}_\gamma(P,B_2)$ for $\lambda \in (0,1)$, unless B_1 and B_2 agree everywhere except where $P(F) = 0$.*

Proof: Logarithmic γ-score is non-negative because $B(F), P(F) \in [0,1]$ for all F so $\log B(F) \leq 0, \gamma(F)P(F) \geq 0$, and $\gamma(F)P(F)\log B(F) \leq 0$.

That $S_\gamma^{\log}(P,B)$ is strictly convex as a function of $\langle\mathbb{B}\rangle$ follows from the strict concavity of $\log x$. Take distinct $B_1, B_2 \in \langle\mathbb{B}\rangle$ and $\lambda \in (0,1)$ and let $B = \lambda B_1 + (1-\lambda)B_2$. Now,

$$\begin{aligned}\gamma(F)P(F)\log(B(F)) &= \gamma(F)P(F)\log(\lambda \cdot B_1(F) + (1-\lambda)B_2(F)) \\ &\geq \gamma(F)P(F)\Big(\lambda \log B_1(F) + (1-\lambda)\log B_2(F)\Big) \\ &= \lambda\gamma(F)P(F)\log B_1(F) + (1-\lambda)\gamma(F)P(F)\log B_2(F)\end{aligned}$$

with equality iff either $P(F) = 0$ or $B_1(F) = B_2(F)$ (since in the latter case $\gamma(F)P(F) > 0$).

Hence,

$$\begin{aligned}S_\gamma^{\log}(P,B) &= -\sum_{F\subseteq\Omega} \gamma(F)P(F)\log B(F) \\ &\leq \lambda S_\gamma^{\log}(P,B_1) + (1-\lambda)S_\gamma^{\log}(P,B_2),\end{aligned}$$

with equality if and only if B_1 and B_2 agree everywhere except possibly where $P(F) = 0$. □

Corollary 84. *For inclusive γ and fixed $P \in \mathbb{P}$, $\arg\inf_{B\in\langle\mathbb{B}\rangle} S_\gamma^{\log}(P,B)$ is unique. For $B' := \arg\inf_{B\in\langle\mathbb{B}\rangle} S_\gamma^{\log}(P,B)$ and for all $F \subseteq \Omega$, we have $B'(F) > 0$ if and only if $P(F) > 0$. Moreover, $B'(\Omega) = 1$ and $B' \in \mathbb{B}$.*

Proof: First of all suppose that there is an $F \subseteq \Omega$ such that $P(F) > 0$ and $B(F) = 0$. Then $S_\gamma^{\log}(P,B) = \infty$. Furthermore, $S_\gamma^{\log}(P,P) < \infty$ for all $P \in \mathbb{P}$. Hence, for $B' \in \arg\inf_{B\in\langle\mathbb{B}\rangle} S_\gamma^{\log}(P,B)$ it holds that $P(F) > 0$ implies $B'(F) > 0$.

Now note that for $P \in \mathbb{P}$ we have $P(\Omega) = 1 - P(\emptyset) = 1$. Furthermore, there are only two partitions $\{\Omega\}$ and $\{\Omega, \emptyset\}$ which contain Ω or $\emptyset$. Minimising $-\gamma(\emptyset)P(\emptyset)\log B'(\emptyset) - \gamma(\Omega)P(\Omega)\log B'(\Omega)$, i.e., $-\gamma(\Omega)\log B'(\Omega)$, subject to the constraint $B'(\emptyset) + B'(\Omega) \leq 1$ is uniquely solved by taking $B'(\Omega) = 1$ and hence $B'(\emptyset) = 0$. Thus, for any B' minimising $S_\gamma^{\log}(P,\cdot)$ it holds that $B'(\emptyset) = 0$ and $B'(\Omega) = 1$. Hence, $B' \in \langle\mathbb{B}\rangle$ is in $\mathbb{B}$.

Now consider a $P \in \mathbb{P}$ such that there is at least one $\emptyset \subset F \subset \Omega$ with $P(F) = 0$. We will show that $B'(F) = 0$ for all $B' \in \arg\inf_{B\in\langle\mathbb{B}\rangle} S_\gamma^{\log}(P,B)$. In the second step we will show that there is a unique infimum B'.

So suppose that the there is a $B' \in \arg\inf_{B\in\langle\mathbb{B}\rangle} S_\gamma^{\log}(P,B)$ such that $B'(F) > 0 = P(F)$. Assume that $\emptyset \subset H \subset \Omega$ is for this B', with respect to subset inclusion, one such largest subset of Ω.

Now define $B'' : \mathscr{P}\Omega \to [0,1]$ by $B''(G) := 0$ for all $G \subseteq H$ and $B''(F) := B'(F)$ otherwise. From $B''(\Omega) = 1, B''(\emptyset) = 0$ we see that $B'' \in \mathbb{B}$; thus $S_\gamma^{\log}(P,B'')$ is well-defined. Since $P \in \mathbb{P}$ we have for all $G \subseteq H$ that $P(H) = P(G) = 0$. Thus, $S_\gamma^{\log}(P,B') = S_\gamma^{\log}(P,B'')$.

Note that since $B' \in \langle\mathbb{B}\rangle$ we have $1 \geq B'(\bar{H}) + B'(H) > B'(\bar{H}) = B''(\bar{H})$. Now define a function $B''' \in \langle\mathbb{B}\rangle$ by

$$\begin{aligned}&B'''(\bar{H}) := 1 \\ &B'''(F) := B''(F) \text{ for all } F \neq \bar{H}.\end{aligned}$$

Since for all $F \subseteq \Omega$, $B''(F) \leq B'''(F)$ and $B''(\bar{H}) < B'''(\bar{H}) = 1$ and $P(\bar{H}) \cdot \gamma(\bar{H}) = 1 \cdot \gamma(\bar{H}) > 0$, we have

$$\begin{aligned} S_\gamma^{\log}(P,B') &= S_\gamma^{\log}(P,B'') \\ &> S_\gamma^{\log}(P,B'''). \end{aligned}$$

We assumed that B' minimises $S_\gamma^{\log}(P,\cdot)$ over $\langle\mathbb{B}\rangle$. Hence, we have a contradiction. We have thus proved that for every $B \in \arg\inf_{B\in\langle\mathbb{B}\rangle} S_\gamma^{\log}(P,B)$, $B(F) = 0$ if and only if $P(F) = 0$. Hence for all $P \in \mathbb{P}$,

$$\arg\inf_{B\in\langle\mathbb{B}\rangle} S_\gamma^{\log}(P,B) = \arg\inf_{\{B\in\langle\mathbb{B}\rangle : P(F)=0 \leftrightarrow B(F)=0\}} S_\gamma^{\log}(P,B). \tag{30}$$

By Proposition 83 we can assume that the right hand side of (30) is a strictly convex optimisation problem on a convex set, which has hence a unique infimum. □

Corollary 85. *$S_\gamma^{\log}$ is strictly $\langle\mathbb{B}\rangle$-proper if and only if $S_\gamma^{\log}$ is strictly $\mathbb{B}$-proper.*

Proof: Assume that $S_\gamma^{\log}$ is strictly $\langle\mathbb{B}\rangle$-proper. Then for all $P \in \mathbb{P}$ we have $P = \arg\inf_{B\in\langle\mathbb{B}\rangle} S_\gamma^{\log}(P,B)$. Since $\mathbb{P} \subset \mathbb{B} \subset \langle\mathbb{B}\rangle$ we hence have $P = \arg\inf_{B\in\mathbb{B}} S_\gamma^{\log}(P,B)$.

For the converse suppose that $S_\gamma^{\log}$ is strictly $\mathbb{B}$-proper, i.e., for all $P \in \mathbb{P}$ we have $P = \arg\inf_{B\in\mathbb{B}} S_\gamma^{\log}(P,B)$. Note that strict propriety implies that γ is inclusive. Corollary 84 implies then that no $B \in \langle\mathbb{B}\rangle \setminus \mathbb{B}$ can minimise $S_\gamma^{\log}(P,B)$. □

Definition 86 (*Symmetric weighting of propositions*). A weighting of propositions γ is *symmetric* if and only if whenever F' can be obtained from F by permuting the ω_i in F, then $\gamma(F') = \gamma(F)$.

Note that γ is symmetric if and only if $|F| = |F'|$ entails $\gamma(F) = \gamma(F')$. For symmetric γ we will sometimes write $\gamma(n)$ for $\gamma(F)$, if $|F| = n$.

Proposition 87. *For inclusive and symmetric γ, $S_\gamma^{\log}$ is strictly $\mathbb{P}$-proper.*

Proof: We have that for all $\omega \in \Omega$, $|\{F \subseteq \Omega : |F| = n,\, \omega \in F\}| = |\{G \subseteq \overline{\{\omega\}} : |G| = n-1\}| = \binom{|\Omega|-1}{n-1}$.

We recall from Example 4 that with $v_n := \binom{|\Omega|-1}{n-1}$ we have

$$\sum_{\substack{F\subseteq\Omega \\ |F|=n}} P(F) = v_n \cdot \sum_{\omega\in\Omega} P(\omega) = v_n.$$

Multiplying the objective function in an optimisation problem by some positive constant does not change where optima obtain. Thus

$$\begin{aligned}\arg\inf_{Q\in\mathbb{P}} -\sum_{\substack{F\subseteq\Omega\\|F|=n}} \gamma(n)P(F)\log Q(F) &= \arg\inf_{Q\in\mathbb{P}} -\sum_{\substack{F\subseteq\Omega\\|F|=n}} \frac{P(F)}{\nu_n}\log Q(F)\\ &= \arg\inf_{Q\in\mathbb{P}} -\sum_{\substack{F\subseteq\Omega\\|F|=n}} \frac{P(F)}{\nu_n}\log(\frac{Q(F)}{\nu_n}\cdot\nu_n)\\ &= \arg\inf_{Q\in\mathbb{P}} -\sum_{\substack{F\subseteq\Omega\\|F|=n}} \frac{P(F)}{\nu_n}\Big(\log\frac{Q(F)}{\nu_n}+\log\nu_n\Big)\\ &= \arg\inf_{Q\in\mathbb{P}} -\sum_{\substack{F\subseteq\Omega\\|F|=n}} \frac{P(F)}{\nu_n}\log\frac{Q(F)}{\nu_n}.\end{aligned}$$

Now note that since $Q,P\in\mathbb{P}$, we have that $\sum_{\substack{F\subseteq\Omega\\|F|=n}} P(F)=\nu_n=\sum_{\substack{F\subseteq\Omega\\|F|=n}} Q(F)$ and hence $\sum_{\substack{F\subseteq\Omega\\|F|=n}} \frac{P(F)}{\nu_n}=1=\sum_{\substack{F\subseteq\Omega\\|F|=n}} \frac{Q(F)}{\nu_n}$. Put $\Psi:=\{F\subseteq\Omega:|F|=n\}$ and let us understand $\frac{P(\cdot)}{\nu_n},\frac{Q(\cdot)}{\nu_n}$ as functions $\frac{P(\cdot)}{\nu_n},\frac{Q(\cdot)}{\nu_n}:\Psi\longrightarrow[0,1]$ with $\sum_{G\in\Psi}\frac{P(G)}{\nu_n}=1=\sum_{G\in\Psi}\frac{Q(G)}{\nu_n}$. It follows that $\frac{P(\cdot)}{\nu_n},\frac{Q(\cdot)}{\nu_n}$ are formally probability functions on Ψ, satisfying certain further conditions which are not relevant in the following. Let $\mathbb{P}_\Psi$ denote the set of probability functions on Ψ and let $\mathbb{P}_\Omega\subseteq\mathbb{P}_\Psi$ be the set of probability functions of the above form $\frac{P(\cdot)}{\nu_n},\frac{Q(\cdot)}{\nu_n}$, where $P,Q\in\mathbb{P}$.

Consider a scoring rule $S(P,B)$ in the standard sense, i.e., expectations over losses are taken with respect to members x of some set X. (At the beginning of §2.4 we considered states $\omega\in\Omega$.) Let $\mathbb{X}$ denote the set of probability functions on the set X. Suppose that S is strictly $\mathbb{X}$-proper. Then for any fixed set $\mathbb{Y}\subseteq\mathbb{X}$ it holds that $\arg\inf_{B\in\mathbb{Y}}S(P,B)=P$ for all $P\in\mathbb{Y}$. It is well-known that the standard logarithmic scoring rule on a given universal set is strictly $\mathbb{X}$-proper. Taking $X=\Psi$, $\mathbb{X}=\mathbb{P}_\Psi$ and $\mathbb{Y}=\mathbb{P}_\Omega$ we obtain for all $\frac{P(\cdot)}{\nu_n}\in\mathbb{P}_\Omega$ that

$$\begin{aligned}\frac{P(\cdot)}{\nu_n} &= \arg\inf_{\frac{Q(\cdot)}{\nu_n}\in\mathbb{P}_\Omega} -\sum_{G\in\Psi}\frac{P(G)}{\nu_n}\log\frac{Q(G)}{\nu_n}\\ &= \arg\inf_{Q\in\mathbb{P}} -\sum_{G\in\Psi}\frac{P(G)}{\nu_n}\log\frac{Q(G)}{\nu_n}.\end{aligned}$$

We thus find:

$$P=\arg\inf_{Q\in\mathbb{P}} -\sum_{\substack{F\subseteq\Omega\\|F|=n}}\gamma(n)P(F)\log Q(F). \tag{31}$$

Since P minimises (31) for every n it also the minimises the sum over all n, and hence

$$P=\arg\inf_{Q\in\mathbb{P}} -\sum_{1\le n\le|\Omega|}\sum_{\substack{F\subseteq\Omega\\|F|=n}}\gamma(F)P(F)\log Q(F)=\arg\inf_{Q\in\mathbb{P}} S_g(P,Q).$$

□

Lemma 88. *If γ is an inclusive weighting of propositions that is equivalent to a weighting of partitions, then $S_\gamma^{\log}$ is strictly $\mathbb{B}$-proper.*

Proof: While this result follows directly from Corollary 18, we shall give another proof which will provide the groundwork for the proof of the next result, Theorem 89.

First we shall fix a $P \in \mathbb{P}$ and observe that the first part of Corollary 84 up to and including (30) still holds with $\mathbb{B}$ substituted for $\langle\mathbb{B}\rangle$. We shall thus concentrate on propositions $F \subset \Omega$ with $P(F) > 0$, since it follows from Corollary 84 that whenever $P(F) = 0$, we must have $B(F) = 0$ and $B(\Omega) = 1$, if $S_\gamma^{\log}(P,B)$ is to be minimised. We thus let $\mathscr{P}^+\Omega := \{\emptyset \subset F \subset \Omega : P(F) > 0\}$ and

$$\begin{aligned}\mathbb{B}^+ := \{B \in \mathbb{B} : 0 < B(F) \leq 1 \text{ for all } F \in \mathscr{P}^+\Omega,\\ B(\Omega) = 1 \text{ and } B(F) = 0 \text{ for all other } F \in \mathscr{P}\Omega \setminus \mathscr{P}^+\Omega\}.\end{aligned}$$

In the following optimisation problem we will thus only consider $B(F)$ to be a variable if $F \in \mathscr{P}^+\Omega$.

We now investigate

$$\arg \inf_{B \in \mathbb{B}^+} S_\gamma^{\log}(P,B). \tag{32}$$

To this end we shall first find for all fixed $t \geq 2$

$$\arg \inf_{\substack{B \in \mathbb{B}^+ \\ B(F) \geq \frac{P(F)}{t} \text{ for all } F \in \mathscr{P}^+\Omega}} S_\gamma^{\log}(P,B). \tag{33}$$

Making this restriction on $B(F)$ allows us to evade any problems which arise from taking the derivative of $\log B(F)$ at $B(F) = 0$ which inevitably arise when we directly apply Karush-Kuhn-Tucker techniques to (32).

With $\Pi' := \{\pi \in \Pi : \pi \neq \{\Omega\},\ \pi \neq \{\Omega, \emptyset\}\}$ we thus need to solve the following optimisation problem:

$$\begin{aligned}
&\text{minimize} && S_\gamma^{\log}(P,B)\\
&\text{subject to} && B(F) \geq \frac{P(F)}{t} > 0 \text{ for } t \geq 2 \text{ and all } F \in \mathscr{P}^+\Omega\\
& && \sum_{\substack{G \in \pi \\ G \in \mathscr{P}^+\Omega}} B(G) \leq 1 \text{ for all } \pi \in \Pi'\\
& && B(\Omega) = 1 \text{ and } B(F) = 0 \text{ for all other } F \in \mathscr{P}\Omega \setminus \mathscr{P}^+\Omega.
\end{aligned}$$

Note that the first and second constraints imply that $0 < B(F) \leq 1$ for all $F \in \mathscr{P}^+\Omega$.

Observe that for $\pi \in \Pi'$ with $G \in \pi$, $|G| \geq 2$ and $P(G) = 0$, there is another partition in Π' which subdivides G and agrees with π everywhere else. These two partitions π, π' will give rise to the exact same constraint on the $F \in \mathscr{P}^+\Omega$. Including the same constraint multiple times does not affect the applicability of the Karush-Kuhn-Tucker techniques. Thus, the solutions of this optimisation problem are the solutions of (33).

With Karush-Kuhn-Tucker techniques in mind we shall define the following function for $B \in \mathbb{B}^+$:

$$\begin{aligned}
Lag(B) &= \overbrace{-\sum_{F \subseteq \Omega} \gamma(F)P(F)\log B(F)}^{S_\gamma^{\log}(P,B)} + \overbrace{\sum_{\pi \in \Pi'} \lambda_\pi \cdot (-1 + \sum_{\substack{G \in \pi \\ G \in \mathscr{P}^+\Omega}} B(G)) + \sum_{F \in \mathscr{P}^+\Omega} \mu_F(\frac{P(F)}{t} - B_F)}^{\text{constraints}}\\
&= -\sum_{F \in \mathscr{P}^+\Omega} \gamma(F)P(F)\log B(F) + \sum_{\pi \in \Pi'} \lambda_\pi \cdot (-1 + \sum_{\substack{G \in \pi \\ G \in \mathscr{P}^+\Omega}} B(G)) + \sum_{F \in \mathscr{P}^+\Omega} \mu_F(\frac{P(F)}{t} - B_F).
\end{aligned}$$

First recall that $B(F) = 0$ iff $P(F) = 0$, thus the first sum is always finite here. Since $B(F) > 0$ for all $F \in \mathscr{P}^+\Omega$ we can take derivatives with respect to the variables $B(F)$. Recalling that $\gamma(F) > 0$ for all $F \subseteq \Omega$ we now find

$$\frac{\partial}{\partial B(F)} Lag(B) = -\gamma(F)\frac{P(F)}{B(F)} + \sum_{\substack{\pi \in \Pi' \\ F \in \pi}} \lambda_\pi - \mu_F \quad \text{for all } F \in \mathscr{P}^+\Omega.$$

Equating these derivatives with zero we obtain

$$\gamma(F)\frac{P(F)}{B(F)} = \sum_{\substack{\pi \in \Pi' \\ F \in \pi}} \lambda_\pi - \mu_F \quad \text{for all } F \in \mathscr{P}^+\Omega. \tag{34}$$

Since γ is by our assumption equivalent to a weighting of partitions, $\gamma(F) = \sum_{\substack{\pi \in \Pi' \\ F \in \pi}} g(\pi)$. Letting $\lambda_\pi := g(\pi), \mu_F := 0$ and $B(F) = P(F)$ for $F \in \mathscr{P}^+\Omega$ solves the set of equations in (34). For $B(F) = P(F)$ when $F \in \mathscr{P}^+\Omega$, we trivially have $\sum_{\substack{G \in \pi \\ G \in \mathscr{P}^+\Omega}} B(G) = 1$ and hence $\lambda_\pi(\sum_{\substack{G \in \pi \\ G \in \mathscr{P}^+\Omega}} B(G) - 1) = 0$. Furthermore, $\mu_F(\frac{P(F)}{t} - B(F)) = 0$ for $F \in \mathscr{P}^+\Omega$.

Thus by the Karush-Kuhn-Tucker Theorem, $B(F) = P(F)$ for $F \in \mathscr{P}^+\Omega$ is a critical point of the optimisation problem (33) for all t and all $P \in \mathbb{P}$ since all constraints are linear.

Note that the constraints $B(\Omega) = 1$, $B(\emptyset) = 0$ and $0 \leq \sum_{F \in \pi} B(F) \leq 1$ for $\pi \in \Pi'$ ensure that B is a member of $\mathbb{B}$ regardless of the actual value of $B(F)$ for $\emptyset \neq F \neq \Omega$. Thus, $B \in \mathbb{B}^+$ if and only if $B(\Omega) = 1$, $B(\emptyset) = 0$, $0 \leq \sum_{F \in \pi} B(F) \leq 1$ for $\pi \in \Pi'$ and $B(F) = 0$ iff $P(F) = 0$. Thus, $\mathbb{B}^+$ is convex. It follows that $B_t^+ := \{B \in \mathbb{B}^+ : B(F) \geq \frac{P(F)}{t}$ for all $F \in \mathscr{P}^+\Omega\}$ is convex for all $t \geq 2$. Since $\mathbb{B}^+$ is the feasible region of (33) the critical point of the convex minimisation problem is the unique minimum.

Letting $t > 0$ tend to 0 we see that $B(F) = P(F)$ for $F \in \mathscr{P}^+\Omega$ is the unique solution of (32).

Thus, any function $B \in \mathbb{B}$ minimizing $S_\gamma^{\log}(P, \cdot)$ has to agree with P on the $F \in \mathscr{P}^+\Omega$. By our introductory remarks it has to hold that $B(\Omega) = 1$ and $B(G) = 0$ for all other $G \subseteq \Omega$. Thus, $B(F) = P(F)$ for all $F \subseteq \Omega$.

We have thus shown that $S_\gamma^{\log}$ is strictly proper. □

Theorem 89. *For inclusive γ with $\gamma(\Omega) \geq \gamma(\emptyset)$, $S_\gamma^{\log}$ is strictly proper if and only if γ is equivalent to a weighting of partitions.*

Proof: From Lemma 88 we have that the existence of the λ_π ensures propriety.

For the converse suppose that $S_\gamma^{\log}$ is strictly $\mathbb{B}$-proper (equivalently, by Corollary 85, strictly proper). By our assumptions we have $\gamma(\Omega) \geq \gamma(\emptyset) > 0$. We can thus put $g(\{\Omega, \emptyset\}) := \gamma(\emptyset)$ and $g(\Omega) := \gamma(\Omega) - \gamma(\emptyset)$. Then $\gamma(\Omega) = g(\{\Omega, \emptyset\}) + g(\Omega) > 0$ and $\gamma(\emptyset) = g(\{\Omega, \emptyset\}) > 0$.

Observe that for all $P \in \mathbb{P}$, for any infimum of the minimisation problem $\arg\inf_{B \in \mathbb{B}} S_\gamma^{\log}(P, B)$ there have to exist multipliers $\lambda_\pi \geq 0$ and $\mu_F \geq 0$ which solve (34) and $\mu_F(\frac{P(F)}{t} - B(F)) = 0$. Now fix a $P \in \mathbb{P}$ such that $P(F) > 0$ for all $\emptyset \subset F \subseteq \Omega$. If $S_\gamma^{\log}$ is strictly $\mathbb{B}$-proper, then the minimisation problem $\arg\inf_{B \in \mathbb{B}} S_\gamma^{\log}(P, B)$ for this P has to be solved uniquely by $B = P$. Thus, strict $\mathbb{B}$-propriety implies that:

$$0 < \gamma(F) = \sum_{\substack{\pi\in\Pi\\ F\in\pi}} \lambda_\pi - \mu_F \quad \text{for all } \emptyset \subset F \subset \Omega \qquad \text{and} \qquad \mu_F \frac{1-t}{t} P(F) = 0 \text{ for all } F \in \mathscr{P}^+\Omega.$$

The latter conditions can only be satisfied if all μ_F vanish. Hence, we obtain the following conditions which necessary have to hold if $S_\gamma^{\log}(P,\cdot)$ is to be uniquely minimised by $B = P$:

$$0 < \gamma(F) = \sum_{\substack{\pi\in\Pi\\ F\in\pi}} \lambda_\pi \quad \text{for all } \emptyset \subset F \subset \Omega.$$

Since all the constraints are inequalities, the corresponding multipliers λ_π have to be greater or equal than zero.

Thus, strict propriety of $S_\gamma^{\log}$ implies the existence of these $\lambda_\pi \geq 0$. This in turn implies that γ is equivalent to a weighting of partitions.

Note that for the purposes of this proof we do not need to investigate what happens if $P \in \mathbb{P}$ is such that there exists a proposition $\emptyset \subset F \subseteq \Omega$ with $P(F) = 0$. □

Note that $\gamma(\Omega) \geq \gamma(\emptyset)$ is not a real restriction. The first component in $S_\gamma^{\log}(\cdot,\cdot)$ is a probability function in the above proof. Thus, $P(\emptyset) = 0$. Hence, $\gamma(\emptyset)P(\emptyset)\log B(\emptyset) = 0$, regardless of $\gamma(\emptyset)$. The particular value of $\gamma(\emptyset)$ is thus irrelevant for strict propriety. So, setting $\gamma(\emptyset) = \gamma(\Omega)$ fulfills the conditions of the Theorem but does not change the value of the γ-score. (The condition is required because if $\gamma(\emptyset) > \gamma(\Omega)$ then, while $S_\gamma^{\log}$ may be strictly proper, it cannot be a weighting of partitions.)

The importance of the condition in Theorem 89 that γ should be equivalent to a weighting of partitions is highlighted in the following:

Example 90. Let $\Omega = \{\omega_1, \omega_2, \omega_3\}$ and $\gamma(1) = \gamma(3) = 1$, and $\gamma(2) = 10$. Now consider $B \in \mathbb{B}$ defined as $B(\emptyset) := 0$, $B(F) := 0.2$ if $|F| = 1$, $B(F) := 0.8$ if $|F| = 2$, and $B(\Omega) := 1$. Then

$$\begin{aligned}
S_\gamma^{\log}(P_=,P_=) &= -\sum_{\omega\in\Omega} P_=(\omega)\log P_=(\omega) \\
&\quad -10\cdot\Big(\sum_{\substack{F\subseteq\Omega\\ |F|=2}} P_=(F)\log P_=(F)\Big) - P_=(\Omega)\cdot\log P_=(\Omega) \\
&= -3\cdot\frac{1}{3}\log\frac{1}{3} - 3\cdot 10\cdot\frac{2}{3}\log\frac{2}{3} \approx 9.2079 \\
S_\gamma^{\log}(P_=,B) &= -\sum_{\omega\in\Omega} P_=(\omega)\log B(\omega) \\
&\quad -10\cdot\Big(\sum_{\substack{F\subseteq\Omega\\ |F|=2}} P_=(F)\log B(F)\Big) - P_=(\Omega)\cdot\log B(\Omega) \\
&= -3\cdot\frac{1}{3}\log 0.2 - 3\cdot 10\cdot\frac{2}{3}\log 0.8 \approx 6.0723
\end{aligned}$$

Thus $S_\gamma^{\log}(P_=,B) < S_\gamma^{\log}(P_=,P_=)$. Hence $S_\gamma^{\log}$ is *not* strictly $\mathbb{B}$-proper, even though γ is inclusive and symmetric. Compare this with Proposition 87, where we proved that positivity and symmetry γ were enough to ensure that $S_\gamma^{\log}$ is strictly $\mathbb{P}$-proper.

Note that strict propriety is exactly what is needed in order to derive Theorem 24, as is apparent from its proof (see also the discussion at the start of Section §2.5). By Theorem 89, only a weighting of propositions that is equivalent to a weighting of partitions can be strictly proper (up to an inconsequential value for $\gamma(\emptyset)$), hence the generalisation of standard entropy and score in the main text, which focusses on weightings of partitions, is essentially the right one for our purposes.

Indeed, adopting a non-strictly proper scoring rule $S_\gamma^{\log}$ may result in Theorem 24 not holding:

Proposition 91. *If $S_\gamma^{\log}$ is not strictly $\mathbb{X}$-proper (with $\mathbb{P} \subseteq \mathbb{X}$), then worst case γ-expected loss minimisation and γ-entropy maximisation are in general achieved by different functions.*

Proof: If S_g is not merely proper, then there is a $P' \in \mathbb{P}$ such that $S_\gamma^{\log}(P', \cdot)$ is not minimised over $\mathbb{X}$ by P'. In particular there is some $Q \in \mathbb{X}$ such that $S_\gamma^{\log}(P', Q) < S_\gamma^{\log}(P', P')$. Suppose that $\mathbb{E} = \{P'\}$. Trivially,

$$\arg\sup_{P \in \mathbb{E}} S_\gamma^{\log}(P, P) = P'.$$

By construction,

$$\begin{aligned}\arg\inf_{Q \in \mathbb{X}} \sup_{P \in \mathbb{E}} S_\gamma^{\log}(P, Q) &= \arg\inf_{Q \in \mathbb{X}} \sup_{P \in \{P'\}} S_\gamma^{\log}(P, Q) \\ &= \arg\inf_{Q \in \mathbb{X}} S_\gamma^{\log}(P', Q) \\ &\not\ni P'.\end{aligned}$$

Thus, the γ-entropy maximiser in $\mathbb{E}$ (here P') is not a function in $\mathbb{X}$ which minimises worst case γ-expected loss.

Finally, consider the case in which $S_\gamma^{\log}$ is merely proper, *i.e.*, there exists a $P' \in \mathbb{P}$ such that $S_\gamma^{\log}(P', \cdot)$ is minimised by both P' and members of a non-empty subset, $\mathbb{Q} \subseteq \mathbb{B} \setminus \{P'\}$. Then, with $\mathbb{E} = \{P'\}$:

$$\arg\inf_{Q \in \mathbb{X}} \sup_{P \in \mathbb{E}} S_\gamma^{\log}(P, Q) = \arg\inf_{Q \in \mathbb{X}} \sup_{P \in \{P'\}} S_\gamma^{\log}(P, Q) = \arg\inf_{Q \in \mathbb{X}} S_\gamma^{\log}(P', Q) = \mathbb{Q} \cup \{P'\}.$$

Thus there is some function other than the γ-entropy maximiser that also minimises γ-score. □

References

Aczél, J. and Daróczy, Z. (1975). *On measures of information and their characterizations.* Academic Press, New York.

Aczel, J. and Pfanzagl, J. (1967). Remarks on the measurement of subjective probability and information. *Metrika*, 11:91–105.

Cover, T. M. and Thomas, J. A. (1991). *Elements of information theory.* John Wiley and Sons, New York.

Csiszàr, I. (2008). Axiomatic characterizations of information measures. *Entropy*, 10(3):261–273.

Dawid, A. P. (1986). Probability forecasting. In Kotz, S. and Johnson, N. L., editors, *Encyclopedia of Statistical Sciences* , volume 7, pages 210–218. Wiley.

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

Jaynes, E. T. (1957). Information theory and statistical mechanics. *The Physical Review*, 106(4):620–630.

Joyce, J. M. (2009). Accuracy and coherence: Prospects for an alethic epistemology of partial belief. In Huber, F. and Schmidt-Petri, C., editors, *Degrees of Belief*, Synthese Library 342. Springer, Netherlands.

Keynes, J. M. (1921). *A treatise on probability*. Macmillan (1948), London.

Keynes, J. M. (1937). The general theory of employment. *The Quarterly Journal of Economics*, 51(2):209–223.

König, H. (1992). A general minimax theorem based on connectedness. *Archiv der Mathematik*, 59:55–64.

Kyburg Jr, H. E. (2003). Are there degrees of belief? *Journal of Applied Logic*, 1:139–149.

McCarthy, J. (1956). Measures of the value of information. *Proceedings of the National Academy of Sciences*, 42(9):654–655.

Paris, J. B. (1994). *The uncertain reasoner's companion*. Cambridge University Press, Cambridge.

Paris, J. B. (1998). Common sense and maximum entropy. *Synthese*, 117:75–93.

Paris, J. B. and Vencovská, A. (1990). A note on the inevitability of maximum entropy. *International Journal of Approximate Reasoning*, 4(3):183–223.

Paris, J. B. and Vencovská, A. (1997). In defense of the maximum entropy inference process. *International Journal of Approximate Reasoning*, 17(1):77–103.

Pettigrew, R. (2011). Epistemic utility arguments for probabilism. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Winter 2011 edition.

Predd, J., Seiringer, R., Lieb, E., Osherson, D., Poor, H., and Kulkarni, S. (2009). Probabilistic coherence and proper scoring rules. *IEEE Transactions on Information Theory*, 55(10):4786–4792.

Ramsey, F. P. (1926). Truth and probability. In Kyburg, H. E. and Smokler, H. E., editors, *Studies in subjective probability*, pages 23–52. Robert E. Krieger Publishing Company, Huntington, New York, second (1980) edition.

Ricceri, B. (2008). Recent advances in minimax theory and applications. In Chinchuluun, A., Pardalos, P., Migdalas, A., and Pitsoulis, L., editors, *Pareto Optimality, Game Theory And Equilibria*, volume 17 of *Optimization and Its Applications*, pages 23–52. Springer.

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

Seidenfeld, T. (1986). Entropy and uncertainty. *Philosophy of Science*, 53(4):467–491.

Shannon, C. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423 and 623–656.

Shuford, E. H., Albert, A., and Massengill, H. E. (1966). Admissible probability measurement procedures. *Psychometrika*, 31(2):125–145.

Smets, P. and Kennes, R. (1994). The transferable belief model. *Artificial Intelligence*, 66(2):191– 234.

Topsøe, F. (1979). Information theoretical optimization techniques. *Kybernetika*, 15:1–27.

Williamson, J. (2010). *In defence of objective Bayesianism*. Oxford University Press, Oxford.

Williamson, J. (2011). An objective Bayesian account of confirmation. In Dieks, D., Gonzalez, W. J., Hartmann, S., Uebel, T., and Weber, M., editors, *Explanation, Prediction, and Confirmation. New Trends and Old Ones Reconsidered*, pages 53–81. Springer, Dordrecht.

Williamson, J. (2013). From Bayesian epistemology to inductive logic. *Journal of Applied Logic*, DOI 10.1016/j.jal.2013.03.006.