Proceedings of Machine Learning Research 103:256–264, 2019 ISIPTA 2019

IP Scoring Rules: Foundations and Applications

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

Abstract

The mathematical foundations of imprecise probability theory (IP) have been in place for 25 years, and IP has proved successful in practice. But IP methods lack rigorous accuracy-centered, philosophical justifications. Traditional Bayesian methods can be justified using epistemic scoring rules, which measure the accuracy of the estimates that they produce. But there has been little work extending these justifications to the IP framework. This paper makes plea for the IP community to embrace this research programme. The plea comes in three parts. Firstly, I outline some initial work developing scoring rules for imprecise probabilities — IP scoring rules — and using them to shore up the philosophical foundations of IP. Secondly, I explain why a range of impossibility results for IP scoring rules should *not* dissuade the IP community from working on the foundations of IP scoring rules. Finally, I highlight one potential applications for IP scoring rules: IP aggregation.

Keywords: IP scoring rules, IP Impossibility Theorems, IP Aggregation

1. Introduction

The mathematical foundations of imprecise probability theory (IP) have been in place since (Walley 1991) [28]. And IP theory has proved successful in practice, e.g., in learning graphical models, classification, action recognition and optical character recognition. But imprecise Bayesian methods lack rigorous accuracy-centered, philosophical justifications. Traditional Bayesian methods can be justified using what are variously known as *epistemic scoring rules*, *epistemic utility functions* or *inaccuracy measures*. Scoring rules measure the accuracy of the estimates that traditional methods produce, which is roughly a matter of how close those estimates are to the actual values of the quantities of interest. Drawing on the work of de Finetti [3] and Savage [22], contemporary Bayesians like Joyce [6, 7], Schervish *et al.* [23] and Pettigrew [18] use scoring rules, together with resources from decision theory, to show that traditional Bayesian methods provide decision-theoretically optimal strategies for securing accurate estimates. This approach has provided compelling justifications for a wide range of traditional Bayesian methods and principles: Probabilism, Conditionalization, the Principle of Indifference and more.

The aim of this paper is twofold: firstly, to outline some initial work extending these justifications to the IP framework; secondly, to respond to some concerns and highlight some potential applications in order to incentivise the IP community to take up this research programme.

2. IP Scoring Rules

Isaac Levi [11, 12, 14] proposed measuring the epistemic value or utility of a state of full belief K at a world w by the following quantity:[1]

$$\mathscr{V}(K,w) = \alpha \cdot \mathbb{E}(K,w) + (1-\alpha) \cdot \mathbb{T}(K,w).$$

The first component, $\mathbb{E}(K,w)$, is the truth-value of belief state K at w (0 if false, 1 if true), *i.e.*, the truth-value of the conjunction of all propositions believed in K. It measures the extent to which K promotes the first of what William James called our "two great commandments as would-be knowers": Avoid error! If K is true, it promotes it fully, *i.e.*, to degree 1 ($\mathbb{E}(K,w) = 1$). If false, it promotes it minimally, *i.e.*, to degree 0 ($\mathbb{E}(K,w) = 0$).

The second component, $\mathbb{T}(K,w)$ is K's degree of informativeness, which is meant to reflect how virtuous K is at w, in terms of "simplicity, explanatory and predictive power, and other allegedly scientific or epistemic" desiderata ([13], p. 83; [15], p. 179). It measures the extent to which K promotes the second of our two great commandments: Seek truth! or as Levi reframes it: Seek Valuable Information! The more informative K is, the closer it comes to promoting it fully, *i.e.*, to degree 1 ($\mathbb{T}(K,w) \approx 1$). The less informative K is, the closer it comes to promoting it minimally, *i.e.*, to degree 0 ($\mathbb{T}(K,w) \approx 0$).

The third component, α, reflects the extent to which you let one or the other of our two principal epistemic aims — Avoid error! Seek Truth! — "color your intellectual life." It measures the respective degree of priority that you give to them.

Putting these components together as follows:

$$\mathscr{V}(K,w) = \alpha \cdot \mathbb{E}(K,w) + (1-\alpha) \cdot \mathbb{T}(K,w)$$

gives us a "summary statistic" that measures the extent to which K succeeds at striking the optimal balance between

1. See [14, §3.1] and [21].

promoting our two principal epistemic aims — Avoid error! Seek Truth! — at w.

We might call Levi's underlying theory of epistemic value the *bipartite theory*:

BTEV. A doxastic state D is more or less epistemically valuable at a world w to the extent that it succeeds at striking the optimal balance between promoting our two principal epistemic aims at w: Avoid error! Seek Truth!

In earlier work [9], I used the bipartite theory of epistemic value to delineate a class of reasonable *scoring rules for imprecise credences* (IP scoring rules). Much like Levi's measure, an IP scoring rule is a function $\mathscr{I}$ which maps IP distributions $\mathscr{C}$ (sets of credence functions) and worlds w to non-negative real numbers, $\mathscr{I}(\mathscr{C},w)$. And like before, $\mathscr{I}(\mathscr{C},w)$ is a summary statistic that measures the extent to which $\mathscr{C}$ succeeds at striking the optimal balance between promoting our two principal epistemic aims at w: Avoid error! Seek Truth! The only difference: Levi's quantity, $\mathscr{V}(K,w)$, captures K's positive epistemic value at w. $\mathscr{I}(\mathscr{C},w)$, on the other hand, captures $\mathscr{C}$'s epistemic *dis*value at w. If $\mathscr{I}(\mathscr{C},w)$ equals 0, then $\mathscr{C}$ is minimally *dis*valuable at w (maximally epistemically valuable). It strikes the *best* possible balance between avoiding error and seeking truth. The larger $\mathscr{I}(\mathscr{C},w)$ is, the less epistemically valuable.

Following Levi, I assumed that $\mathscr{I}$ takes the following form:

$$\mathscr{I}_\alpha(\mathscr{C},w) = \alpha \cdot \mathscr{E}(\mathscr{C},w) + (1-\alpha) \cdot \mathscr{T}(\mathscr{C},w)$$

The first component, $\mathscr{E}(\mathscr{C},w)$, is meant to measure the extent to which $\mathscr{C}$ promotes the first of our two principal epistemic aims: Avoid error! An IP distribution *avoids error* to the extent that it *leaves open* accurate credences functions. In [9], I proposed measuring how well $\mathscr{C}$ leaves open accurate credence functions at w by:

$$\mathscr{E}(\mathscr{C},w) = \min_{c \in \mathscr{C}} I(c,w)$$

where I is any (precise) strictly proper scoring rule.

The second component, $\mathscr{T}(\mathscr{C},w)$, is meant to measure the extent to which $\mathscr{C}$ promotes the second of our two principal epistemic aims: Seek truth! An IP distribution *pins down the truth* to the extent that it *rules out* inaccurate credence functions. I proposed measuring how well $\mathscr{C}$ rules out inaccurate credence functions at w by:

$$\mathscr{T}(\mathscr{C},w) = \max_{c \in \mathscr{C}} I(c,w)$$

Once more, α measures the degree to which you prioritise avoiding error over seeking truth, or vice versa. So the proposed class of IP scoring rules take the following form:

$$\mathscr{I}_\alpha(\mathscr{C},w) = \alpha \cdot \min_{c \in \mathscr{C}} I(c,w) + (1-\alpha) \cdot \max_{c \in \mathscr{C}} I(c,w)$$

Clearly IP scoring rules of this form are overly simplistic. Nevertheless, they are still useful for highlighting some of the foundational purposes that IP scoring rules might serve. For example, in [9], I show that for any "conservative" IP scoring rule of this form ($\alpha \geq 0.5$) and any conservative Hurwicz decision rule — which recommends deciding between priors by a weighted average of their worst-case score $\mathscr{I}^-_{\mathscr{C}}$ and best-case score $\mathscr{I}^+_{\mathscr{C}}$, *i.e.* $\beta\mathscr{I}^-_{\mathscr{C}} + (1-\beta)\mathscr{I}^+_{\mathscr{C}}$ (conservative if $\beta \geq 0.5$) — then some imprecise (non-singleton) $\mathscr{C}$ is optimal. The upshot, roughly, is that imprecise priors do a better job than precise priors at balancing the risk of worst-case catastrophe and the chance best-case spoils. Any agent who adopts precise priors, rather than imprecise ones, goes wrong by gambling with the epistemic utility of her doxastic state in too risky a fashion. Precise priors represent an overly risky epistemic bet, according to conservative IP scoring rules and Hurwicz criteria.

This argument helps to illuminate our most basic epistemic reasons for preferring IP distributions to precise distributions. But it also makes a range of simplifying assumptions about which factors make IP distributions epistemically valuable, how those factors determine overall epistemic value, and how to measure such value. In addition, it justifies only one plank of imprecise probability theory.

Over the coming years, I hope to provide fully adequate epistemic justifications for all of imprecise probability theory. This will shore up the philosophical foundations of IP. My goal for the remainder of this paper is (i) to explain why a range of *impossibility results* for IP scoring rules should *not* dissuade the IP community from working on the foundations of IP scoring rules, and (ii) highlight one cluster of applications for IP scoring rules.

3. Impossibility Theorems for IP Scoring Rules

Seidenfeld *et al.* [25], Schoenfield [24], and Mayo-Wilson and Wheeler [16] provide a range of impossibility theorems that seem to threaten the viability of IP scoring rules. These theorems show that continuous IP scoring rules must fail to have certain *prima facie* desirable properties. In particular, every such scoring rule $\mathscr{I}$ renders some IP distribution $\mathscr{C}$ dominated (guaranteed to be less valuable according to $\mathscr{I}$ than some other IP distribution $\mathscr{C}^*$). In this case, the IP scoring rule in question fails to satisfy a constraint known as *admissibility*. On the face of it, this seems like a troubling result. Every major result in epistemic utility theory — a research programme which provides accuracy-centred justifications for Bayesian norms using scoring rules — relies on the assumption that precise scoring rules render probability distributions non-dominated. So it seems natural to expect reasonable IP scoring rules to render every

IP distribution (set of probabilities) non-dominated. The impossibility theorems for IP scoring rules show that no continuous IP scoring rule can do so. This leads various authors to conclude that IP methods cannot be motivated by purely accuracy-centred considerations ([24, p. 14], [16, p. 15]).

This however is the wrong moral to draw. We should *want* IP scoring rules to violate admissibility. Different IP distributions are better at securing certain types of competing alethic goods than others. Very roughly, more imprecise distributions (more inclusive sets of probabilities) do better at avoiding error. This comes at a cost however. More imprecise distributions are less able to positively pin down the truth, even in their respective best cases. In contrast, more precise distributions have better best-case proximity to truth. They positively pin down the truth better in the best case. But this comes at a cost. It opens them up to greater error.

Different IP scoring rules reflect different ways of prioritising these competing alethic goods (error-avoidance and proximity to truth). So any IP scoring rule *should* render certain IP distributions dominated because certain IP distributions are *guaranteed* to strike a better balance between these competing alethic goods than others. IP scoring rules *must* violate admissibility to reflect this fact.

Moreover, the fact that IP scoring rules violate admissibility does not threaten their usefulness for providing accuracy-centred justifications for IP methods. Historically, epistemic utility theorists have aimed to provide accuracy-centred justifications for traditional Bayesian methods and principles by showing that there is *some* reasonable measure of epistemic value (precise scoring rule) and *some* plausible decision rule that together deliver the same prescriptions as the method or principle in question. This does not require any *single* scoring rule or any *single* decision rule to feature in the justification of every prima facie plausible method or principle. Likewise, we do not need any single IP scoring rule and any single decision rule to feature in the justification of every IP method or principle. It might well be that certain ways of prioritising competing alethic goods justify certain IP methods or principles (perhaps ones that recommend more imprecise distributions), while other ways of prioritising them justify other methods or principles (perhaps ones that recommend more precise distributions).

On the face of it, impossibility theorems also threaten the usefulness of IP scoring rules for elicitation. If an IP scoring rule $\mathscr{I}$ renders one IP distribution $\mathscr{B}$ dominated by another $\mathscr{C}$ — so that $\mathscr{C}$ is guaranteed to receive a smaller penalty than $\mathscr{B}$ according to $\mathscr{I}$ — then $\mathscr{I}$ cannot be used to incentivise a rational agent whose opinions are captured by $\mathscr{B}$ to report $\mathscr{B}$ as her IP distribution. She can always do better by reporting $\mathscr{C}$. It seems natural to conclude that since every reasonable IP scoring renders some IP distribution dominated, IP scoring rules are simply not useful tools for elicitation.

But again this is the wrong moral to draw. IP scoring rules *can* be useful for elicitation if one uses supplemental elicitation techniques to tailor the IP scoring rule to the agent in question. In fact, both precise scoring rules and IP scoring rules need to be tailored in this way. As Savage [22, §10.4] was well aware, strictly proper (precise) scoring rules do not in general incentivise rational agents to report their own precise distribution. For example, an expected utility maximizer with £3 in her pocket, log utilities, and credences $\langle 0.1, 0.2, 0.7\rangle$ on a 3-cell partition will report previsions $\langle 0.136, 0.243, 0.621\rangle$ if previsions are penalised by the Brier score. So supplemental elicitation techniques are needed to gather information about the agent's utility function and tailor an appropriate scoring rule for the agent. Similarly, we might use standard IP elicitation techniques [1, ch. 15], together with tools for eliciting probabilities in more complex decision frameworks (*e.g.*, the weighted expected utility framework) to tailor IP scoring rules that are useful for elicitation on an agent-by-agent basis.

The upshot is this: impossibility results for IP scoring rules pose no serious threat to the use of IP scoring rules in providing value-based justifications for norms governing imprecise credence.

Before moving on, it will prove instructive to explore precisely *how* the IP scoring rules in §2 violate admissibility, and why it is *desirable* that they do so. In the course of our exploration, we will restrict our attention to the simplest possible case. In particular, we will focus an expressively weak IP model—lower and upper probabilities—and restrict our attention to a single proposition X. So, for example, we will consider agents who are between 10% and 40% confident in X, *i.e.*, whose lower and upper probabilities for X are 0.1 and 0.4, respectively. In addition, we will assume that our IP scoring rule $\mathscr{I}_\alpha$ is generated by a strictly proper (precise) scoring rule I (*e.g.*, the Brier score, spherical score, log score, etc.).

Recall, our IP scoring rules take the form

$$\mathscr{I}_\alpha(\mathscr{C}, w) = \alpha \cdot \mathscr{E}(\mathscr{C}, w) + (1-\alpha) \cdot \mathscr{T}(\mathscr{C}, w)$$

In the case at hand this boils down to

$$\mathscr{I}_\alpha([x,y], 0) = \alpha \cdot I(x, 0) + (1-\alpha) \cdot I(y, 0)$$

and

$$\mathscr{I}_\alpha([x,y], 1) = \alpha \cdot I(y, 1) + (1-\alpha) \cdot I(x, 1)$$

When we examine the class of IP scoring rules of this form, what we see is that for each value of α, a different range of lower probability models are *non-dominated*. For $\alpha \leq 0.5$, for example, all and only the precise lower probability functions (*i.e.* lower probability = upper probability) are non-dominated. So relative to $\mathscr{I}_{0.5}$, it is permissible to have

any precise probability x between 0 and 1 (inclusive) for X, but no properly imprecise probability.

To have a single way to visualise precise and imprecise probabilities, represent a precise probability x as a degenerate interval $[x,x]$. Plot intervals $[x,y]$ as pairs $\langle x,y\rangle$ in the unit square $[0,1]^2$. So precise probabilities x between 0 and 1 (inclusive) correspond to points on the line $y = x$.

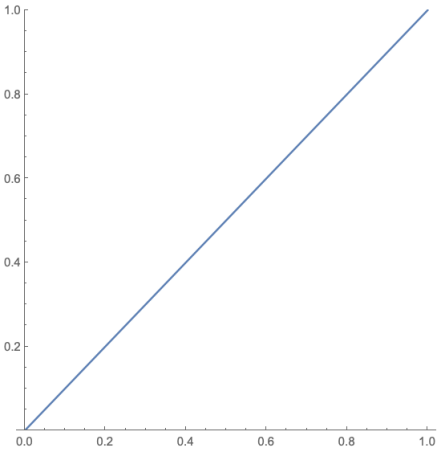


Fig 1: Non-dominated credences rel. to $\mathscr{I}_{0.5}$, viz., the precise credences $[x,x]$ with $0 \leq x \leq 1$, plotted as points $\langle x,x\rangle$ in $[0,1]^2$.

For $\alpha = 0.7$, in contrast, the non-dominated lower/upper probability models (intervals) are all *imprecise* (save for 0 and 1). The intervals that are non-dominated relative to $\mathscr{I}_{0.7}$ live on the following curve:

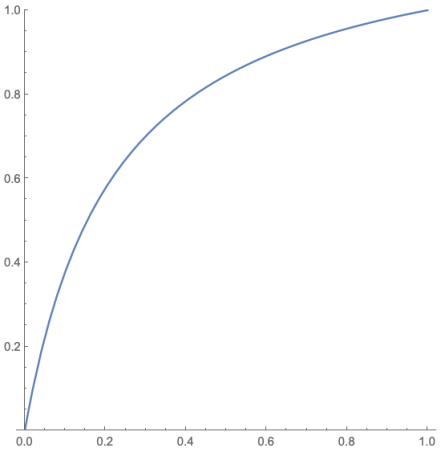


Fig 2: Non-dominated credal states rel. to $\mathscr{I}_{0.7}$

For example, the following intervals (lower and upper probabilities) are all permissible (non-dominated): $[0.1, 0.376923]$, $[0.2, 0.576471]$, $[0.3, 0.7]$, $[0.4, 0.784]$, $[0.5, 0.844828]$, $[0.6, 0.890909]$, $[0.7, 0.927027]$, $[0.8, 0.956098]$, $[0.9, 0.98]$. These are plotted as points on the curve in figure 2, e.g., $\langle 0.1, 0.376923\rangle$, $\langle 0.2, 0.576471\rangle$, etc.

For each distinct α and α^* (≥ 0.5), the sets of non-dominated intervals relative to $\mathscr{I}_\alpha$ and $\mathscr{I}_{\alpha^*}$ are disjoint (save for $[0,0]$ and $[1,1]$ which are always non-dominated). What's more, these sets of non-dominated intervals — each one given by a different curve through the unit square — partition the space of possible intervals. So for every interval, $[x,y]$, there's *some* IP scoring rule that renders it non-dominated. These curves partition the space of possible interval-valued credences as follows:

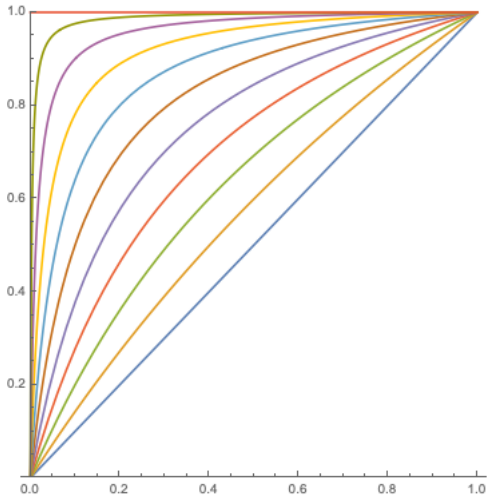


Fig 3: Non-dominated credal states rel. to $\mathscr{I}_\alpha$ for a range of α values between 0.5 and 1.

The picture that emerges is something like this. How you prioritise avoiding error and positively pinning down the truth, respectively (captured by α), fixes the stock of IP distributions (in the case at hand: intervals, or lower/upper probability functions) available to you. Each such stock of credal states has a similar structure. No one has fewer or greater degrees of freedom than any other. If your IP scoring rule $\mathscr{I}_\alpha$ is maximally liberal ($\alpha \leq 0.5$), then you will always adopt a precise probability for X. For example, you might adopt a probability of 0.5 for X. But if my IP scoring rule is more conservative, e.g., $\alpha = 0.7$, then I will have a different—and more imprecise—stock of intervals available to me. Where you adopt probability 0.5, for example, I will adopt the interval $[0.3, 0.7]$. Where you adopt probability 0.4, I might for example adopt credence $[0.209547, 0.590719]$. (Both determine the same epistemic utility ratio if our $\mathscr{I}_\alpha$'s are generated by the Brier score, *viz.*, $\mathscr{I}([x,y],0)/\mathscr{I}([x,y],1) = 0.4444$.)

We see different lower and upper probabilities as appropriate responses to the same evidence not because we disagree about the strength of the evidence, whether or not it is ambiguous, etc., but rather because we take different attitudes toward the comparative importance of avoiding error and pinning down the truth, and different types of lower/upper probabilities (intervals) do a better job at one or the other.

The fact that each IP scoring rule yields a single curve of non-dominated intervals, and that these curves partition the space of possible intervals as per above, will prove crucial

in characterising our new rule for aggregating imprecise probabilities.

4. Current Approaches to Aggregating Imprecise Credences

IP aggregation principles tell us how to construct an IP distribution that reflects a group's opinions when group members individually have imprecise opinions. Stewart and Quintana [26], Elkin and Wheeler [4], Nau [17] and Kriegler *et al.* [10] specify various IP aggregation principles. These authors take a broadly axiomatic approach. On this approach, you first pin down desirable properties of an aggregation principle (e.g., IP aggregation principles should preserve unanimous independence judgments). Then you identify the methods of aggregation that have these properties.

It will prove instructive to consider the virtues and vices of some of these IP aggregation principles, and of the axiomatic approach to IP aggregation more generally. Due to space constraints, we will consider only two such principles: *convex IP pooling* and *linear pooling of lower probabilities*.

4.1. Convex IP Pooling (Stewart and Quintana)

Convex IP pooling takes any finite collection of IP distributions, $\mathscr{C}_1,\ldots,\mathscr{C}_n$, and delivers the convex hull of their union, $conv\{\cup_i \mathscr{C}_i\}$ as the aggregate IP distribution. In the simple case of interest to us, convex IP pooling takes any finite collection of intervals (lower and upper probabilities) for X, $[a_1,b_1],\ldots,[a_n,b_n]$, and delivers the convex hull of their union, $conv\{\cup_i [a_i,b_i]\}$ as the aggregate interval.

Convex IP pooling has a number of *prima facie* desirable properties. For example, Stewart and Quintana [26], as well as Elkin and Wheeler [4], both find some version of the *preservation of irrelevance principle* plausible.

> *Preservation of Irrelevance (PIE)*: If X is irrelevant to Y according to each of $\mathscr{C}_1,\ldots,\mathscr{C}_n$, then X should be irrelevant to Y according to their aggregate $\mathscr{P}$.

Convex IP pooling satisfies PIE, at least in one sense. It is a familiar refrain amongst imprecise Bayesians that univocal concepts in the traditional Bayesian framework fracture into families of related concepts in the IP framework. For example, Couso *et al.* [5] and Cozman [2] survey a number of concepts of irrelevance/independence in the IP framework that collapse into a single concept in the traditional Bayesian framework. Here are two versions of PIE, which Stewart and Quintana [26] call *confirmational irrelevance preservation* (CIP) and *stochastic independence preservation* (SIP):

> **CIP**: If $\{c(X|Y)|c\in\mathscr{C}_i\}=\{c(X)|c\in\mathscr{C}_i\}$ for all i, then $\{p(X|Y)|p\in\mathscr{P}\}=\{p(X)|p\in\mathscr{P}\}$.

> **SIP**: If $c(X|Y)=c(X)$ for all $c\in\mathscr{C}_i$ and all i, then $p(X|Y)=p(X)$ for all $p\in\mathscr{P}$.

Convex IP pooling satisfies CIP but violates SIP. In addition to preserving irrelevance/independencies, in the sense specified by CIP, convex IP pooling also *commutes with conditionalization*. To make this more precise, let F be an IP aggregation function. So $F(\mathscr{C}_1,\ldots,\mathscr{C}_n)$ is the aggregate of IP distributions $\mathscr{C}_1,\ldots,\mathscr{C}_n$.

Let $\mathscr{C}^E$ be the result of pointwise conditionalizing $\mathscr{C}$ on E, *i.e.*,

$$\mathscr{C}^E=\{c(\cdot|E)|c\in\mathscr{C}\}$$

Let $F^E(\mathscr{C}_1,\ldots,\mathscr{C}_n)$ be the result of pointwise conditionalizing $F(\mathscr{C}_1,\ldots,\mathscr{C}_n)$ on E. Then F commutes with conditionalization (CC) just in case it satisfies the following:

> **CC**: The result of pointwise conditionalizing $\mathscr{C}_1,\ldots,\mathscr{C}_n$ on E and then aggregating is the same as the result of aggregating $\mathscr{C}_1,\ldots,\mathscr{C}_n$ and then pointwise conditionalizing on E, *i.e.*,
>
> $$F^E(\mathscr{C}_1,\ldots,\mathscr{C}_n)=F(\mathscr{C}_1^E,\ldots,\mathscr{C}_n^E)$$

If an aggregation principle violates CC, then it seems to saddle us with inconsistent verdicts about what the group should think and do over time. On the face of it, $F^E(\mathscr{C}_1,\ldots,\mathscr{C}_n)$ and $F(\mathscr{C}_1^E,\ldots,\mathscr{C}_n^E)$ both capture, in some sense, what the group's state of opinion should be after learning E. So if $F^E(\mathscr{C}_1,\ldots,\mathscr{C}_n)\neq F(\mathscr{C}_1^E,\ldots,\mathscr{C}_n^E)$, then there is no single, consistent answer to what the group should think post-learning. Moreover, Buckak *et al.* [8] argue that if an aggregation principle violates CC, then it opens the group up to a diachronic Dutch book. So it is a point in favour of convex IP pooling that it satisfies CC.

While these properties of convex IP pooling are *prima facie* desirable, they are not dispositive. For example, Pettigrew [19] rejects PIE for precise aggregation principles on the basis that structural properties of precise probability distributions such as $p(X|Y)=p(X)$ don't encode *judgments* at all. So there is no reason to expect reasonable aggregation principles to preserve such properties, even when all group members share them. There is reason to be wary of CC as well. IP aggregation principles strike a compromise between individual IP distributions. In doing so, they wash out certain information contained in individual IP distributions. This provides some reason to update-and-then-aggregate rather than aggregate-and-then-update. Updating-and-then-aggregating makes more information available for the purpose of processing new evidence E, which plausibly results in a better-informed group posterior.

Convex IP pooling also has some positively troubling features. It captures the *consensus* amongst individual group members, rather than striking a *compromise* between their individual IP distributions. But consensus is a high bar. IP

distributions that reflect consensus will often be rather uninformative. Though such distributions undoubtedly have their place—*e.g.*, in representing common ground amongst group members—the fact that they wash out such a great deal of information distributed across the group makes them ill-suited to (i) inform future research (as the aggregates that feature in meta-analyses of medical research do), and (ii) serve as an input to decision-theory (as the aggregates that feature in summaries for policy makers do).

More worryingly, *convex IP pooling yields dominated aggregates*. More carefully, convex IP pooling often aggregates my interval (lower and upper probability) for X, $[x,y]$, and your interval, $[a,b]$, in a way that we both agree is epistemically defective in virtue of being *epistemic utility dominated*. Suppose for example that my interval for X is $[0.1,0.376923]$, and your interval is $[0.4,0.784]$. Then convex IP pooling delivers $[0.1,0.784]$ as our aggregate. But $[0.1,0.784]$ is defective from our respective perspectives. Recall, each IP scoring rule yields a single curve of non-dominated intervals, and these curves partition the space of possible intervals. $\mathscr{I}_{0.7}$ is the only IP scoring rule that renders my interval of $[0.1,0.376923]$ for X non-dominated. In that sense, it is "my" IP scoring rule. Similarly, $\mathscr{I}_{0.7}$ is the only IP scoring rule that renders your interval of $[0.4,0.784]$ for X non-dominated. So we both agree, in a sense, that the appropriate way to balance our two principle epistemic aims of avoiding error and pinning down the truth is captured by $\mathscr{I}_{0.7}$. But—and here is the problem for convex IP pooling—$[0.1,0.784]$ is *dominated* relative to our shared IP scoring rule $\mathscr{I}_{0.7}$. We both agree that a much more informative interval for X, *viz.*, $[0.256418,0.652473]$ is *guaranteed* to be better than $[0.1,0.784]$:

$$\mathscr{I}_{0.7}([0.256418,0.652473],0) = 0.173742 < 0.191397 = \mathscr{I}_{0.7}([0.1,0.784],0)$$

and

$$\mathscr{I}_{0.7}([0.256418,0.652473],1) = 0.250417 < 0.275659 = \mathscr{I}_{0.7}([0.1,0.784],1)$$

4.2. Linear Pooling of Lower Probabilities (IP Linear Pooling)

IP linear pooling takes any collection $[x_1,y_1]\dots[x_n,y_n]$ of intervals for X and outputs any interval of the form

$$\left[\sum_i \lambda_i x_i, \sum_i \lambda_i y_i\right]$$

with $\sum_i \lambda_i = 1$ as a permissible aggregate.

Linear pooling is an attractive method for aggregating precise probability distributions. Carl Wagner [27] shows that linear pooling is the only aggregation method that (i) makes the group probability for X depend only on individual probabilities for X and (ii) outputs a group credence of zero whenever all individuals have zero credence. Richard Pettigrew shows that linear pooling alone preserves unanimous judgments about the comparative preferability of actions [20]. In addition, linear pooling is supported by an expected accuracy argument [19]. On top of this, IP linear pooling is fairly robust against outliers (unlike convex IP pooling), and yields informative intervals even when individual intervals are relatively wide [10, pp. 5045-46].

But linear pooling also faces a number of well-known objections. For example, linear pooling fails to preserve irrelevancies/independences, and also fails to commute with conditionalization. In addition, various desirable features of linear pooling of precise probability distributions do not carry over to IP linear pooling. For example, Pettigrew's expected accuracy argument for linear pooling has no analogue in the IP setting. More importantly, *IP linear pooling yields dominated aggregates* just as convex IP pooling does. The same is true of Elkin and Wheeler [4] and Nau's [17] proposed IP aggregation principles.

5. Epistemic Utility Based Aggregation

We will now outline and defend a family of novel methods for aggregating lower and upper probabilities. To do so, we will need a few tools. Firstly, if you have lower and upper probabilities given by $[a,b]$ for X, then we will call $\mathscr{I}_\alpha$ with

$$\alpha = \frac{-b+ab+\sqrt{ab-a^2b-ab^2+a^2b^2}}{a-b}$$

"*your IP scoring rule*." The reason is this: this is the *unique* IP scoring rule (of the form outlined in §2) that renders your interval *non-dominated*. For any $\beta \neq \alpha$, $\mathscr{I}_\beta$ renders the interval $[a,b]$ for X dominated in the sense that there's some other interval $[x,y]$ such that

$$\mathscr{I}_\beta([x,y],0) < \mathscr{I}_\beta([a,b],0)$$

and

$$\mathscr{I}_\beta([x,y],1) < \mathscr{I}_\beta([a,b],1)$$

Next, if $\mathscr{I}_\alpha$ is my IP scoring rule and $\mathscr{I}_\beta$ is your IP scoring rule, then any $\mathscr{I}_\gamma$ with $\alpha \leq \gamma \leq \beta$ is a *compromise* between our respective scoring rules. The rationale is roughly this. α captures how epistemically liberal/conservative I am. And β captures how epistemically liberal/conservative you are. Any γ between our respective degrees of epistemic liberalness/conservativity determines an IP scoring rule $\mathscr{I}_\gamma$ that captures a compromise between our respective epistemic values. More generally, if n individuals have epistemic utility functions $\mathscr{I}_{\alpha_1},\dots,\mathscr{I}_{\alpha_n}$, then any $\mathscr{I}_\gamma$ with $\min_i \alpha_i \leq \gamma \leq \max_i \alpha_i$ is a compromise.

We will present epistemic utility based aggregation in steps, starting with the simplest case and working up to more complex ones.

5.1. The Simple Case

In the simplest possible case, n individuals have intervals $[a_1,b_1],\ldots,[a_n,b_n]$ for X, and they all share an IP scoring rule $\mathscr{I}_\alpha$. In this case, epistemic utility based aggregation recommends the following.

> **EU Aggregation (Unique IP Scoring Rule)**: If n individuals have intervals $[a_1,b_1],\ldots,[a_n,b_n]$ for X, and they all share an epistemic utility function $\mathscr{I}_\alpha$, then any reasonable aggregate must take the following form:
>
> $$\left[x, \frac{\alpha^2 x}{1-2\alpha+\alpha^2-x+2\alpha x}\right]$$
>
> where $\min_i a_i \leq x \leq \max_i a_i$.

If our n individuals share an IP scoring rule, $\mathscr{I}_\alpha$, then $[a_1,b_1],\ldots,[a_n,b_n]$ all sit on the same curve of non-dominated intervals. The proposed IP aggregation principle really says two things. Firstly, it says that reasonable aggregates must sit on the same curve, and so are non-dominated relative to the shared IP scoring rule $\mathscr{I}_\alpha$. Secondly, it says that reasonable aggregates must sit between the two most extreme intervals, *i.e.*, the ones that are closest to $[0,0]$ (closest to certain that X is false) and $[1,1]$ (closest to certain that X is true).

Consider, for example, individuals with the following intervals for X: $[0.1,0.376923]$, $[0.2,0.576471]$ and $[0.5,0.844828]$. These individuals share the IP scoring rule $\mathscr{I}_{0.7}$. EU Aggregation demands that reasonable aggregates sit on the segment of the following curve that intersects with the light blue region:

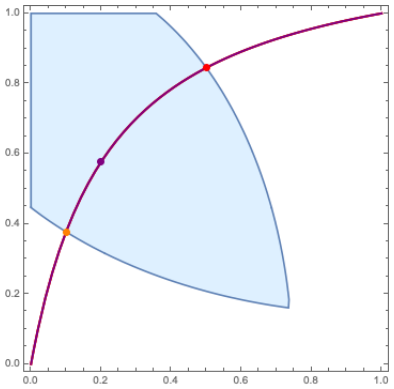


Fig 4: Aggregates of $[0.1,0.376923]$, $[0.2,0.576471]$ *and* $[0.5,0.844828]$.

This simple form of epistemic utility based aggregation is justified by an epistemic utility argument. Here it is:

1. **Measure of Epistemic Value**: The epistemic value or utility of lower/upper probabilities $[a,b]$ for X at a world w is measured by

$$\mathscr{I}_\alpha([a,b],w) = \alpha \cdot \min_{x\in[a,b]} I(x,w) + (1-\alpha)\cdot \max_{x\in[a,b]} I(x,w)$$

2. **Dominance**: If n individuals have credences $[a_1,b_1],\ldots,[a_n,b_n]$ for X and a shared epistemic utility function $\mathscr{I}_\alpha$, then any reasonable aggregate $[x,y]$ should be non-dominated relative to $\mathscr{I}_\alpha$.

3. **No Worse Than the Worst**: Any reasonable aggregate $[x,y]$ should never be worse than the epistemic utility of the worst individual in the group, *i.e.*, $\mathscr{I}_\alpha([x,y],w) \leq \max_i \mathscr{I}_\alpha([a_i,b_i],w)$.

4. **Theorem**: **Dominance** and **No Worse Than the Worst** are satisfied by all and only the aggregates permitted by **EU Aggregation (Unique IP Scoring Rule)**.

C. **EU Aggregation (Unique IP Scoring Rule)**.

Premise 2 seems non-negotiable. If all individuals share an IP scoring rule—which recall reflects the epistemic values that rationalise their individual lower and upper probabilities—and they agree that $[x,y]$ is dominated by $[x',y']$ relative to that scoring rule, then $[x,y]$ is a suboptimal compromise between their individual (non-dominated) lower and upper probabilities. Premise 3 is less secure and proposed only as a working hypothesis. The basic thought is this: if we want EU Aggregation to underpin some type of "wisdom of the crowds" result (something like: aggregates have more epistemic utility than the average utilities of individuals), then something like **No Worse Than the Worst** seems desirable. Of course, it *could* turn out that aggregation procedures which allow aggregates to *sometimes* do worse than the worst end up reliably outperforming procedures that *always* require aggregates to be at least as good as the worst. But barring some positive reason to think that we *must* trade off between having a not-too-low floor (guaranteed to be at least as good as the worst individual) and having high average epistemic utility (or perhaps high objective expected epistemic utility), it seems we ought to shoot for both. The former might even help to secure the latter.

5.2. The Slightly Less Simple Case

In the slightly less simple case case, n individuals have intervals $[a_1,b_1],\ldots,[a_n,b_n]$ for X, but do not share an IP scoring rule. In future work, I hope to characterise the class of aggregates that make the following true:

> **EU Aggregation (No Unique IP Scoring Rule)**: If n individuals have interval-valued credences $[a_1,b_1],\ldots,[a_n,b_n]$ for X, and they have (possibly distinct) IP scoring rules $\mathscr{I}_{\alpha_i}$, then for any reasonable aggregate $[x,y]$, the following should be true: there is some compromise $\mathscr{I}_\beta$ between the $\mathscr{I}_{\alpha_i}$'s that makes $[x,y]$ satisfy **Dominance** and **No Worse Than the Worst**.

Though I have no such general characterisation to offer here, it is straightforward enough to specify the relevant class in any given case. Consider, for example, two individuals with the following intervals for X: $[0.238, 0.393]$ and $[0.668, 0.966]$. The first interval is rendered non-dominated by $\mathscr{I}_{0.59}$, the second by $\mathscr{I}_{0.791}$. So **EU Aggregation (No Unique IP Scoring Rule)** demands that reasonable aggregates sit in the following region, which is bounded by the curves of non-dominated intervals associated with $\mathscr{I}_{0.59}$ and $\mathscr{I}_{0.791}$.

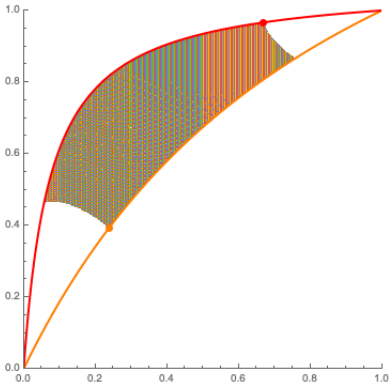
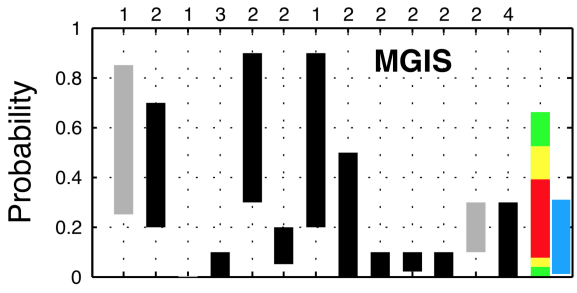


Fig 5: Aggregates of $[0.238, 0.393]$ *and* $[0.668, 0.966]$.

To close, I will apply EU aggregation to a data set from Kriegler *et al*. [10] and briefly outline some virtues of EU aggregation. Kriegler *et al*. elicited lower and upper probabilities for the occurrence of a number of climate events from 43 scientists. We will only consider one particular event: the melting of the Greenland ice sheet (MGIS) assuming the most extreme temperature corridor (global mean temperature capped at 2° increase through 2200). Kriegler *et al*. aggregate individual intervals using IP linear pooling. Individual expert intervals are pictured in black; IP linear pools using a range of weightings are pictured in red, yellow and green.



To apply epistemic utility based aggregation, I used a simple uniform weighting of individual IP scoring rules $\mathscr{I}_{\alpha_i}$ to arrive at a compromise IP scoring rule $\mathscr{I}_{\beta}$. I also used a uniform weighting of individual estimates to determine an aggregate on the curve of non-dominated intervals relative to $\mathscr{I}_{\beta}$. Clearly, a more sophisticated method for setting individual expert weights is desirable (*e.g.*, based individual self-assessments of expertise, or cross-assessment of expertise). The resulting aggregate is given by $[0.0153442, 0.318378]$ (pictured in blue).

Epistemic utility based aggregation has a number of epistemic and practical virtues. Firstly, and most importantly, it yields non-dominated aggregates. Also, as illustrated by the application to the Kriegler *et al*. data set, EU based aggregates strike a genuine *compromise* between individual lower and upper probabilities. They do not simply represent *consensus* amongst the group. As such, EU based aggregates are informative enough to serve as and input to decision theory and inform future research. Finally, like IP linear pooling, epistemic utility based aggregation is relatively robust against outliers.

References

[1] T. Augustin, F. P. Coolen, G. de Cooman, , and M. C. Troffaes, editors. *Introduction to Imprecise Probabilities*. John Wiley and Sons, New York, 2014.

[2] Fabio Cozman. Sets of probability distributions, independence, and convexity. *Synthese*, 186:577–600, 2012.

[3] Bruno de Finetti. *Theory of Probability. A Critical Introductory Treatment*. John Wiley & Sons, 1974.

[4] L. Elkin and G. Wheeler. Resolving peer disagreements through imprecise probabilities. *Nous*, 52(2): 260–278, 2016.

[5] Peter Walley Ines Cuoso, Serafin Moral. A survey of concepts of independence for imprecise probabilities. *Risk, Decision and Policy*, 5(2):165–181, 2000.

[6] J Joyce. A nonpragmatic vindication of probabilism. *Philosophy of Science*, 65(4):575–603, 1998.

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

[8] Lara Buchak J.S. Russell, John Hawthorne. Groupthink. *Philosophical Studies*, 172:1287–1309, 2015.

[9] Jason Konek. Epistemic conservativity and imprecise credence. *Philosophy and Phenomenological Research*, 2019.

[10] Elmar Kriegler. Imprecise probability assessment of tipping points in the climate system. *Proceedings of the National Academy of Sciences*, 106(13):5041–5046, 2009.

[11] Isaac Levi. *Gambling with Truth*. Knopf, New York, 1967.

[12] Isaac Levi. *Decisions and Revisions*. Cambridge University Press, Cambridge, 1984.

[13] Isaac Levi. *The Fixation of Belief and its Undoing: Changing Beliefs through Inquiry*. Cambridge University Press, Cambridge, 1991.

[14] Isaac Levi. *Mild Contraction: Evaluating Loss of Information due to Loss of Belief*. Oxford University Press, Oxford, 2004.

[15] Isaac Levi. *Pragmatism and Inquiry: Selected Essays*. Oxford University Press, Oxford, 2012.

[16] Conor Mayo-Wilson and Gregory Wheeler. Accuracy and imprecision: A mildly immodest proposal. *Philosophy and Phenomenological Research*, 2015.

[17] Robert Nau. The aggregation of imprecise probabilities. *Journal of Statistical Planning and Inference*, 105(1):265–282, 2002.

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

[19] Richard Pettigrew. On the accuracy of group credences. In Tamar Szabo Gendler and John Hawthorne, editors, *Oxford Studies in Epistemology*, volume 6. Oxford University Press, 2016.

[20] Richard Pettigrew. A pragmatic characterisation of linear pooling. *Ms.*, 2018.

[21] Hans Rott. The Value of Truth and the Value of Information. *Knowledge and inquiry: Essays on the pragmatism of Isaac Levi*, page 179, 2006.

[22] L.J. Savage. *The Foundations of Statistics*. Dover, New York, 1972.

[23] Mark Schervish, Teddy Seidenfeld, and Jay Kadane. Proper scoring rules, dominated forecasts, and coherence. *Decision Analysis*, 6(4):202–221, 2009.

[24] Miriam Schoenfield. The accuracy and rationality of imprecise credences. *Nous*, 2015.

[25] T Seidenfeld, M J Schervish, and J B Kadane. Forecasting with imprecise probabilities. *International Journal of Approximate Reasoning*, 53:1248–1261, 2012.

[26] R.T. Stewart and I.O. Quintana. Probabilistic opinion pooling with imprecise probabilities. *J Philos Logic*, 47:17–45., 2018.

[27] Carl Wagner. Allocation, lehrer models, and the consensus of probabilities. *Theory and Decision*, 14: 207–220, 1982.

[28] Peter Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, New York, 1991.