Second-order Probabilism: Expressive Power and Accuracy

Rafal Urbaniak and Marcello Di Bello

2023-09-06

Table of contents

DISCLAIMER: This is a draft of work in progress, please do not cite or distribute without permission.

Abstract.

need to write one when done

Introduction

Precise probabilism (PP) has it that a rational agent's (RA) uncertainty is to be represented as a single probability measure. The view has been criticized on the ground that RA's degrees of belief are not appropriately evidence-responsive, especially when evidence is scant. Accordingly, an alternative view—imprecise probabilism (IP)—has been proposed, on which RA's uncertainty is to be represented by a set of probability measures, rather than a unique one.

Unfortunately, this view runs into problems as well. (1) It still does not seem to be sufficiently evidence-responsive, (2) it is claimed to get certain comparative probability judgments wrong, (3) it seems to be unable to model learning when the starting point is complete lack of information, and (4) notoriously there exist no inaccuracy measure of an imprecise credal stance if the measure is to satisfy certain straightforward formal conditions.

Think about including synergy

The main claim of this paper is that the way forward is to use higher-order probabilities to represen't RA's uncertainty in the relevant cases. The key idea is that uncertainty is not a single-dimensional thing to be mapped on a single one-dimensional scale like a real line and that it's the whole shape of the whole distribution over parameter values that should be taken under consideration. This guiding idea can be used to resolve many problems and philosophical puzzles raised in the debate between PP and IP. Moreover, Bayesian probabilistic programming already provides a fairly reliable implementation framework of this approach.

add structure description

Precise vs. imprecise probabilisms

Precise probabilism

Precise probabilism (PP) holds that a rational agent's uncertainty about a hypothesis is to be represented as a single, precise probability measure. This is an elegant and simple theory. But representing our uncertainty about a proposition in terms of a single, precise probability runs into a number of difficulties. Precise probabilism—arguably—fails to capture an important dimension of how our fallible beliefs reflect the evidence we have (or have not) obtained. A couple of stylized examples should make the point clear. For the sake of simplicity, we will use examples featuring coins.

> **No evidence v. fair coin** You are about to toss a coin, but have no evidence whatsoever about its bias. You are completely ignorant. Compare this to the situation in which you know, based on overwhelming evidence, that the coin is fair.

On precise probabilism, both scenarios are represented by assigning a probability of .5 to the outcome *heads*. If you are completely ignorant, the principle of insufficient evidence suggests that you assign .5 to both outcomes. Similarly, if you know for sure the coin is fair, assigning .5 seems the best way to quantify the uncertainty about the outcome. The agent's evidence in the two scenario is quite different, but precise probabilities fail to capture this difference.

> **Learning from ignorance** You toss a coin with unknown bias. You toss it 10 times and observe *heads* 5 times. Suppose you toss it further and observe 50 *heads* in 100 tosses.

Since the coin initially had unknown bias, you should presumably assign a probability of .5 to both outcomes if you stick with PP. After the 10 tosses, you end up again with an estimate of .5. You must have learned something, but whatever that is, it is not modeled by precise probabilities. When you toss the coin 100 times and observe 50 heads, you learn something new as well. But your precise probability assessment will again be .5.

These examples suggest that precise probabilism is not appropriately responsive to evidence when it comes to representing what RA justifiedly believes of has learned. It ends up assigning the same probability in situations in which one's evidence is quite different: when no evidence is available about the coin's bias; when there is little evidence that the coin is fair (say, after only 10 tosses); and when there is strong evidence that the coin is fair (say, after 100 tosses). The general problem is, precise probability captures the value around which your uncertainty should be centered, but fails to capture how centered it should be given the evidence.[1]

Precise probabilism, it has been argued, fails also to account for cases in which an agent remains undecided even after some additional evidence has been obtained. Imagine RA doesn't know what the bias of the coin is, which PP represents as $\mathsf{P}(()H) = .5$. Then she learns that the bias towards heads has been slightly increased by .001 (in the philosophical literature, this is called *sweetening*. Intuitively, this might still leave RA equally undecided when it comes to betting on H. that would've been fair even if the actual chance of H was .5 and not .001. The same sweetening, however, should make RA bet on H if their original lack of information was in fact correctly captured as a precise credence.

Imprecise probabilism

What if we give up the assumption that probability assignments should be precise? Imprecise probabilism (IP) holds that an agent's credal stance towards a hypothesis is to be represented by means of a *set of probability measures*, typically called a representor $\mathbb{P}$, rather than a single measure P. The representor should include all and only those probability measures which are compatible with the evidence. For instance, if an agent knows that the coin is fair, their credal state would be represented by the singleton set $\{\mathsf{P}\}$, where P is a probability measure which assigns .5 to *heads*. If, on the other hand, the agent knows nothing about the coin's bias, their credal state would be represented by the set of all probabilistic measures, since none of them is excluded by the available evidence. Note that the set of probability measures does not represent admissible options that the agent could legitimately pick from. Rather, the agent's credal state is essentially imprecise and should be represented by means of the entire set of probability measures.[2]

[1]In fact, analogous problems arise even if we do not start with complete lack of evidence; if RA initially weakly believes that the coin is .6 biased towards heads, as she might still learn more, by confirming her belief by tossing the coin repeatedly and observing, say, 60 heads in 100 tosses—but this improvement is not mirrored in the precise probability she will assign to heads.

[2]For the development of imprecise probabilism, see Keynes (1921); Levi (1974); Gärdenfors & Sahlin (1982); Kaplan (1968); Joyce (2005); Fraassen (2006); Sturgeon (2008); Walley (1991). Bradley (2019) is a good source of further references. Imprecise probabilism shares some similarities with what we might call **interval probabilism** (Kyburg, 1961; **kyburg2001uncertain?**). On interval probabilism, precise probabilities are replaced by intervals of probabilities. On imprecise probabilism, instead, precise probabilities are replaced by sets of probabilities. This makes imprecise probabilism more general, since the probabilities of a proposition in the representor set do not have to form a closed interval. In what follows, we will ignore interval probabilism, as intervals do not contain probabilistic information sufficient to guide reasoning with multiple items of evidence.

Imprecise probabilism, at least *prima facie*, offers a straightforward picture of learning from evidence, that is a natural extension of the classical Bayesian approach. When faced with new evidence E between time t_0 and t_1, the representor set should be updated point-wise, running the standard Bayesian updating on each probability measure in the representor:

$$\mathbb{P}_{t_1} = \{\mathsf{P}_{t_1} | \exists\, \mathsf{P}_{t_0} \in \mathbb{P}_{t_0}\ \forall H\ \left[\mathsf{P}_{t_1}(H) = \mathsf{P}_{t_0}(H|E)\right]\}.$$

The hope is that, if we start with a range of probabilities that is not extremely wide, point-wise learning will behave appropriately. For instance, if we start with a prior probability of *heads* equal to .4 or .6, then those measure should be updated to something closer to .5 once we learn that a given coin has already been tossed ten times with the observed number of heads equal 5 (call this evidence E). This would mean that if the initial range of values was $[.4, .6]$ the posterior range of values should be more narrow.

But even this seemingly straightforward piece of reasoning is hard to model without using densities. For to calculate $\mathsf{P}(\mathsf{bias} = k|E)$ we need to calculate $\mathsf{P}(E|\mathsf{bias} = k)\mathsf{P}(\mathsf{bias} = k)$ and divide it by $\mathsf{P}(E) = \mathsf{P}(E|\mathsf{bias} = k)\mathsf{P}(\mathsf{bias} = \mathsf{k}) + \mathsf{P}(E|\mathsf{bias} \neq k)\mathsf{P}(\mathsf{bias} \neq k)$. The tricky part is obtaining $\mathsf{P}(\mathsf{bias} = k)$ or $\mathsf{P}(\mathsf{bias} \neq k)$ in a principled manner without explicitly going second-order, without estimating the parameter value and without using beta distributions.

The situation is even more difficult if we start with complete lack of knowledge, as imprecise probabilism runs into the problem of **belief inertia** (Levi, 1980). Say you start tossing a coin knowing nothing about its bias. The range of possibilities is $[0, 1]$. After a few tosses, if you observed at least one tail and one heads, you can exclude the measures assigning 0 or 1 to *heads*. But what else have you learned? If you are to update your representor set point-wise, you will end up with the same representor set. Consequently, the edges of your resulting interval will remain the same. In the end, it is not clear how you are supposed to learn anything if you start from complete ignorance.

Here's another example from Rinard (2013). Either all the marbles in the urn are green (H_1), or exactly one tenth of the marbles are green (H_2). Your initial credence is complete uncertainty with interval $[0, 1]$ associated with each hypothesis. Then you learn that a marble drawn at random from the urn is green (E). After conditionalizing each function in your representor on this evidence, you end up with the the same spread of values for H_1 that you had before learning E, and no matter how many marbles are sampled from the urn and found to be green.

Some downplay the problem of belief inertia. They insist that vacuous priors should not be used and that imprecise probabilism gives the right results when the priors are non-vacuous. After all, if you started with knowing truly nothing, then perhaps it is right to conclude that you will never learn anything. Another strategy is to say that, in a state of complete ignorance, a special updating rule should be deployed.[3] But no matter what we think about belief inertia, other problems plague imprecise probabilism. Three problems are particularly pressing.

[3] Lee (2017) suggests the rule of *credal set replacement* that recommends that upon receiving evidence the agent should drop measures rendered implausible, and add all non-extreme plausible probability measures. This, however, is tricky. One needs a separate account of what makes a distribution plausible or not, as well as a principled account of why one should use a separate special update rule when starting with complete ignorance.

One problem is that **imprecise probabilism fails to capture intuitions we have about evidence and uncertainty in a number of scenarios.** Consider this example:

> **Even v. uneven bias:** You have two coins and you know, for sure, that the probability of getting heads is .4, if you toss one coin, and .6, if you toss the other coin. But you do not know which is which. You pick one of the two at random and toss it. Contrast this with an uneven case. You have four coins and you know that three of them have bias .4 and one of them has bias .6. You pick a coin at random and plan to toss it. You should be three times more confident that the probability of getting heads is .4. rather than .6.

The first situation can be easily represented by imprecise probabilism. The representor would contain two probability measures, one that assigns .4. and the other that assigns .6 to the hypothesis 'this coin lands heads'. But imprecise probabilism cannot represent the second situation, at least not without moving to higher-order probabilities or assigning probabilities to chance hypotheses, in which case it is no longer clear whether the object-level imprecision does any heavy lifting.[4]

Second, besides descriptive inadequacy, imprecise probabilism fases a foundational problem. It arises when we attempt to measure the accuracy of a representor set of probability measures. Workable *scoring rules* exist for measuring the accuracy of a single, precise credence function, such as the Brier score. These rules measure the distance between one's credence function (or probability measure) and the actual value. A requirement of scoring rules is that they be *proper*: any agent will score their own credence function to be more accurate than every other credence function. After all, if an agent thought a different credence was more accurate, they should switch to it. Proper scoring rules are then used to formulate accuracy-based arguments for precise probabilism. These arguments show (roughly) that, if your precise credence follows the axioms of probability theory, no other credence is going to be more accurate than yours whatever the facts are. Can the same be done for imprecise probabilism? It seems not. Impossibility theorems demonstrate that **no proper scoring rules are available for representor sets**. So, as many have noted, the prospects for an accuracy-based argument for imprecise probabilism look dim (Campbell-Moore, 2020; Mayo-Wilson & Wheeler, 2016; Schoenfield, 2017; Seidenfeld, Schervish, & Kadane, 2012). Moreover, as shown by Schoenfield (2017), if an accuracy measure satisfies certain plausible formal constraints, it will never strictly recommend an imprecise stance, as for any imprecise stance there will be a precise one with at least the same accuracy.

The third problem with imprecise probabilism is that, **degenerate cases aside, it is hard to make sense of the notion of an IP agent learning that a probabilistic measure is incompatible with the evidence.** Recall that the probability measures allowed in a representor set are supposed to be only those compatible with

[4] Other scenarios can be constructed in which imprecise probabilism fails to capture distinctive intuitions about evidence and uncertainty; see, for example, (Rinard, 2013). Suppose you know of two urns, GREEN and MYSTERY. You are certain GREEN contains only green marbles, but have no information about MYSTERY. A marble will be drawn at random from each. You should be certain that the marble drawn from GREEN will be green (G), and you should be more confident about this than about the proposition that the marble from MYSTERY will be green (M). In line with how lack of information is to be represented on IP, for each $r \in [0,1]$ your representor contains a P with $\mathsf{P}(M) = r$. But then, it also contains one with $\mathsf{P}(M) = 1$. This means that it is not the case that for any probability measure P in your representor, $\mathsf{P}(G) > \mathsf{P}(M)$, that is, it is not the case that RA is more confident of G than of M. This is highly counter-intuitive.

the agent's evidence. The idea is that thanks to this feature, imprecise credal stances are evidence-responsive in a way precise probabilistic stances are not. But how, exactly, does the evidence exclude probability measures?

This is not a mathematical question: mathematically (**bradley2012scientific?**), evidential constraints are easy to model. They can take the form, for example, of the *evidence of chances* $\{\mathsf{P}(X) = x\}$ or $\mathsf{P}(X) \in [x, y]$, or *structural constraints* such as "X and Y are independent" or "X is more likely than Y." While it is clear that these constraints are something that an agent can come to accept if offered such information by an expert to which the agent completely defers, it is not trivial to explain how non-testimonial evidence can result in such constraints for an epistemic agent that functions as IP proposes.

Most of the examples in the literature start with the assumption that the agent is told by a believable source that the chances are such-and-such, or that the experimental set-up is such that the agent knows that such and such structural constraint is satisfied. But, besides ideal circumstances, it is unclear how an agent could come to accept such structural constraints upon observation. The chain of testimonial evidence has to end somewhere.

Admittedly, there are straightforward degenerate cases: if you see the outcome of a coin toss to be heads, you reject the measure with $\mathsf{P}(H) = 0$, and similarly for tails. Another class of cases might arise if you are randomly drawing objects from a finite set where the real frequencies are already known, because this finite set has been inspected. But such extreme cases aside, what else? Mere consistency constraint wouldn't get the agent very far in the game of excluding probability measures, as way too many probability measures are strictly speaking still consistent with the observations for evidence to result in epistemic progress.^[Bradley suggests that "statistical evidence might inform [evidential] constraints [...and that evidence] of causes might inform structural constraints" [125-126]. This, however, is not a clear account of how exactly this should proceed. One suggestion might be that once a statistical significance threshold is selected, a given set of observations with a selection of background modeling assumptions yields a credible interval. But this is to admit that to reach such constraints, we already have to start with a second-order approach, and drop information about the densities, focusing only on the intervals obtained with fixed margins of errors. But as we will be insisting, if you have the information about densities to start with, there is no clear advantage to going imprecise instead, and there are multiple problems associated with this move. Moreover, such moves require a choice of an error margin, which is extra-epistemic, and it is not clear what advantage there is to use extra-epistemic considerations of this sort to drop information contained in densities.[5]

Higher-order probabilism

There is, however, a view in the neighborhood that fares better: a higher-order perspective. In fact, some of the comments by the proponents of imprecise probabilism tend to go in this direction. For instance, Bradley compares the measures in a representor to committee members, each voting on a particular issue, say the true bias of a coin. As they acquire more evidence, the committee members will often converge on a specific chance hypothesis. He writes (**bradley2012scientific?**):

[5]Relatedly, in forensic evidence evaluation even scholars who disagree about the value of going higher-order agree that interval reporting is problematic, as the choice of a limit or uncertainty level is rather arbitrary (**Taroni2015Dismissal?**; **Sjerps2015Uncertainty?**).

…the committee members are "bunching up". Whatever measure you put over the set of probability functions—whatever "second order probability" you use—the "mass" of this measure gets more and more concentrated around the true chance hypothesis'.

Note, however, that such bunching up cannot be modeled by imprecise probabilism alone.[6]

In a similar vein, Joyce (2005), in a paper defending imprecise probabilism, attempts to explicate something that imprecise probabilism was advertised to handle better than precise probabilism: weight of evidence. But in fact, the explication uses a density over chance hypotheses to account for the notion of evidential weight and conceptualizes the weight of evidence as an increase of concentration of smaller subsets of chance hypotheses, without any reference to representors in the explication of the notion of weight.

#(we will get back to his explication when we discuss weight of evidence).

The idea that one should use higher-order probabilities has also been suggested by critics of imprecise probabilism. For example, Carr (2020) argues that sometimes evidence requires uncertainty about what credences to have. Carr, however, does not articulate this suggestion more fully, does not develop it formally, and does not explain how her approach would fare against the difficulties affecting precise ad imprecise probabilism. This is the key goal of this paper.

The underlying idea of the higher-order approach we propose is that **uncertainty is not a single-dimensional thing to be mapped on a single one-dimensional scale such as a real line. It is the whole shape of the whole distribution over parameter values that should be taken under consideration.**[7] From this perspective, when an agent is asked about their credal stance towards X, they can refuse to summarize it in terms of a point value $\mathsf{P}(X)$. They can instead express their credal stance in terms of a probability (density) distribution f_x treating $\mathsf{P}(X)$ as a random variable. To be sure, an agent's credal state toward X could sometimes be usefully represented by the expectation, especially when the agent is quite confident about the probability of a given proposition.

Generally, expectation is defined as $\int_0^1 x f(x)\, dx$–in the context of our approach here, we can think of x as the objectively appropriate/justified degree of belief in a given proposition, and of f as the density representing the agent's uncertainty about x. Perhaps, such an expectation can be used as the precise, object-level credence in the proposition itself, where f is the probability density over possible object-level probability values. But this need not always be the case. If the probability density f is not sufficiently concentrated around a single value, a one-point summary might fail to do justice to the nuances of the agent's credal state. This approach lines up with common practice in Bayesian statistics, where the primary role of uncertainty representation is assigned to the whole distribution. Summaries such as the mean, mode standard deviation, mean absolute deviation, or highest posterior density intervals are only succinct ways for representing the uncertainty of a given scenario.

For example, consider again the scenario in which the agent knows that the bias of

[6]Bradley seems to be aware of that, which would explain the use of scare quotes: when he talks about the option of using second-order probabilities in decision theory, he insists that 'there is no justification for saying that there is more of your representor here or there.' ~[p.~195]

[7]Bradley admits this much (**bradley2012scientific?**), and so does Konek (Konek, 2013, p. 59). For instance, Konek disagrees with: (1) X is more probable than Y just in case $p(X) > p(Y)$, (2) D positively supports H if $p_D(H) > p(H)$, or (3) A is preferable to B just in case the expected utility of A w.r.t. p is larger than that of B.