Second-order Probability, Accuracy and Weight of Evidence

Rafal Urbaniak and Marcello Di Bello

November 24, 2022

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DISCLAIMER: This is a draft of work in progress, please do not cite or distribute without permission.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

1 Introduction

this is what a comment about what will go into book looks like; will be globally supressed when generating latex for the journal paper, don't worry about deleting them.

A defendant in a criminal case may face multiple items of incriminating evidence whose strength can at least sometimes be assessed using probabilities. For example, consider a murder case in which the police recover trace evidence that matches the defendant. Hair found at the crime scene matches the defendant's hair (call this evidence hair). In addition, the defendant owns a dog whose fur matches the dog fur found in a carpet wrapped around one of the bodies (call this evidence dog).[1] The two matches suggest that the defendant (and the defendant's dog) must be the source of the crime traces (call this hypothesis source). But how strong is this evidence, really? What are the fact-finders to make of it?

The standard story among legal probabilists goes something like this. To evaluate the strength of the two items of match evidence, we must find the value of the likelihood ratio:

$$\frac{\mathsf{P}(\mathsf{dog} \wedge \mathsf{hair}|\mathsf{source})}{\mathsf{P}(\mathsf{dog} \wedge \mathsf{hair}|\neg\mathsf{source})}$$

For simplicity, the numerator can be equated to one. To fill in the denominator, an expert provides the relevant random match probabilities. Suppose the expert testifies that the probability of a random person's hair matching the reference sample is about 0.0253, and the probability of a random dog's hair matching the reference sample happens to be about the same, 0.0256.[2] Presumably, the two matches are independent lines of evidence. In other words, their random match probabilities must be independent of each other conditional on the source hypothesis. Then, to evaluate the overall impact of the evidence on the source hypothesis, you calculate:

$$\begin{aligned}\mathsf{P}(\mathsf{dog} \wedge \mathsf{hair}|\neg\mathsf{source}) &= \mathsf{P}(\mathsf{dog}|\neg\mathsf{source}) \times \mathsf{P}(\mathsf{hair}|\neg\mathsf{source}) \\ &= 0.0252613 \times 0.025641 = 6.4772626 \times 10^{-4}\end{aligned}$$

This is a very low number. Two such random matches would be quite a coincidence. Following our advice from Chapter 5, the expert facilitates your understanding of how this low number should be interpreted. They show you how the items of match evidence change the probability of the source hypothesis given a range of possible priors (Figure 1). The posterior of .99 is reached as soon as the prior is higher than 0.061.[3] While perhaps not sufficient for outright belief in the source hypothesis, the evidence seems extremely strong: a minor additional piece of evidence could make the case against the defendant overwhelming.

[1] The hair evidence and the dog fur evidence are stylized after two items of evidence in the notorious 1981 Wayne Williams case (Deadman, 1984b, 1984a).

[2] Probabilities have been slightly but not unrealistically modified to be closer to each other in order to make a conceptual point. The original probabilities were 1/100 for the dog fur, and 29/1148 for Wayne Williams' hair. We modified the actual reported probabilities slightly to emphasize the point that we will elaborate further on: the same first-order probabilities, even when they sound precise, may come with different degrees of second-order uncertainty.

[3] These calculations assume that the probability of a match if the suspect and the suspect's dog are the sources is one.

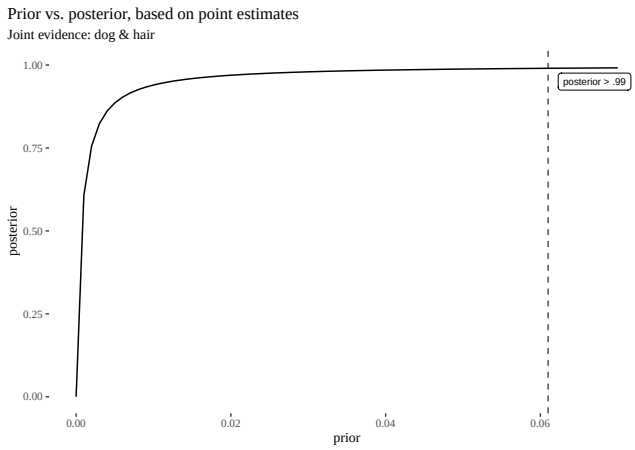

Figure 1: Impact of dog fur and human hair evidence on the prior, point estimates.

Unfortunately, this analysis leaves out something crucial. You reflect on what you have been told and ask the expert: how can you know the random match probabilities with such precision? Shouldn't we also be mindful of the uncertainty that may affect these numbers? The expert agrees, and tells you that in fact the random match probability for the hair evidence is based on 29 matches found in a database of size 1148, while the random match probability for the dog evidence is based on finding two matches in a reference database of size 78.

The expert's answer makes apparent that the precise random match probabilities do not tell the whole story. Perhaps, the information about sample sizes is good enough and now you know how to use the evidence properly.[4] But if you are like most human beings, you can't. What to do, then?

added this bit to draw attention to this aspect of the Taroni debate, to come back to this

You ask the expert for guidance: what are reasonable ranges of the random match probabilities? What are the worst-case and best-case scenarios? The expert responds with 99% credible intervals—specifically, starting with uniform priors, the ranges of the random match probabilities are (.015,.037) for hair evidence and (.002, .103) for fur evidence.[5] With this information, you redo your calculations using the upper bounds of the two intervals: .037 and .103. The rationale for choosing the upper bounds is that these numbers result in random match probabilities that are most favorable to the defendant. Your new calculation yields the following:

$$\mathsf{P}(\mathsf{dog} \wedge \mathsf{hair} | \neg \mathsf{source}) = .037 \times .103 = .003811.$$

This number is around 5.88 times greater than the original estimate. Now the prior probability of the source hypothesis needs to be higher than 0.274 for the posterior probability to be above .99 (Figure 2). So you are no longer convinced that the two items of match evidence are strongly incriminating.

[4] This is what, effectively, CITE TARONI seem to suggest when they insist the fact-finders should be simply given point estimates and information about the study set-up, such as sample size. As will transpire, we disagree.

[5] Roughly, the 99% credible interval is the narrowest interval to which the expert thinks the true parameter belongs with probability .99. For a discussion of what credible intervals are, how they differ from confidence intervals, and why confidence intervals should not be used, see Chapter 3.

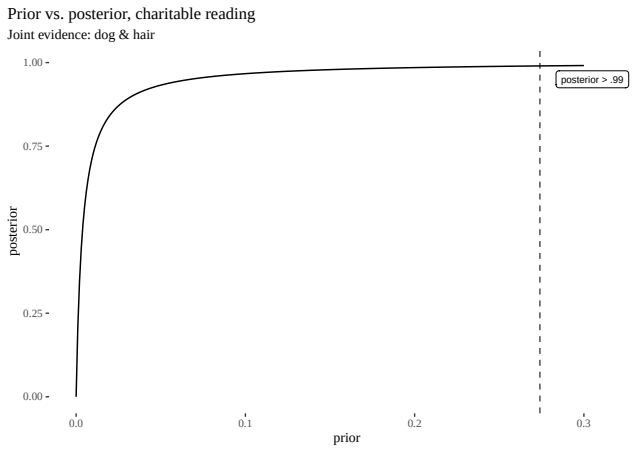

Figure 2: Impact of dog fur and human hair evidence on the prior, charitable reading.

This result is puzzling. Are the two items of match evidence strongly incriminating evidence (as you initially thought) or somewhat weaker (as the new calculation suggests)? For one thing, using precise random match probabilities might be too unfavorable toward the defendant. On the other hand, your new assessment of the evidence based on the upper bounds might be too *favorable* toward them. Is there a middle way that avoids overestimating and underestimating the strength of the evidence?

To see what this middle path looks like, we should reconsider the calculations you just did. You made an important blunder: you assumed that because the worst-case probability for one event is x and the worst-case probability for another independent event is y, the worst-case probability for their conjunction is xy. But this conclusion does not follow if the margin of error (credible interval) is fixed. The intuitive reason is simple: just because the probability of an extreme (or larger absolute) value x for one variable X is .01, and so it is for the value y of another independent variable Y, it does not follow that the probability that those two independent variables take values x and y simultaneously is the same. This probability is actually much smaller. The interval presentation instead of doing us good led us into error.

In general, it is impossible to calculate the credible interval for the joint distribution based solely on the individual credible intervals corresponding to the individual events. We need additional information: the distributions that were used to calculate the intervals for the probabilities of the individual events. In our example, if you additionally knew, for instance, that the expert used beta distributions (as, arguably, they should in this context), you could in principle calculate the 99% credible interval for the joint distribution. It usually will not be the same as whatever the results of multiplication of individual interval edges, and it is unlikely that a human fact-finder would be able to correctly run such calculations in their head even if they knew the functional form of the distributions used. [6] So providing the fact-finder with individual intervals, even if further information about the distributions is provided, might easily mislead.[7]

Can you google to see if there is any such study?

As it turns out, given the reported sample sizes, the 99% credible interval for the probability $\mathsf{P}(\mathsf{dog} \wedge \mathsf{hair} | \neg \mathsf{source})$ is $(0.000023, 0.002760)$.

The upper bound of this interval would then require the prior probability of the source hypothesis to be above .215 for the posterior to be above .99. On this interpretation, the two items of match evidence are still not quite as strong as you initially thought, but stronger than what your second calculation indicated.

the fn was repetitive, compare to fn 5

Still, the interval approach—even the corrected version just outlined—suffers from a more general problem. Working with intervals might be useful if the underlying distributions are fairly symmetrical. But in our case, they might not be. For instance, Figure 3 depicts beta densities for dog fur and human hair, together with sampling-approximated density for the joint evidence. The distribution for the joint evidence is not symmetric. If you were only informed about the edges of the interval, you would be

[6] Also, in principle, in more complex contexts, we need further information about how the items of evidence are related if we cannot take them to be independent.

[7] Investigation of the extent to which the individual interval presentation is misleading would be an interesting psychological study.

oblivious to the fact that the most likely value (and the bulk of the distribution, really) does not simply lie in the middle between the edges. Just because the parameter lies in an interval with some posterior probability, it does not mean that the ranges near the edges of the interval are equally likely—the bulk of the density might very well be closer to one of the edges. Therefore, only relying on the edges can lead one to either overestimate or underestimate the probabilities at play. This also means that—following our advice on how to illustrate the impact of evidence on prior probabilities—a better representation of the dependence of the posterior on the prior should comprise multiple possible sampled lines whose density mirrors the density around the probability of the evidence (Figure 4).

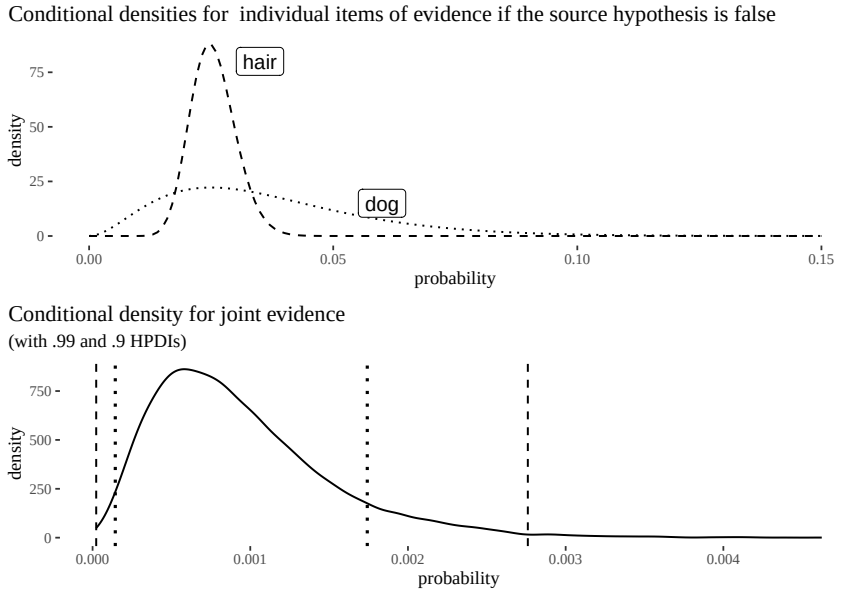

Figure 3: Beta densities for individual items of evidence and the resulting joint density with .99 and .9 highest posterior density intervals, assuming the sample sizes as discussed and independence, with uniform priors.

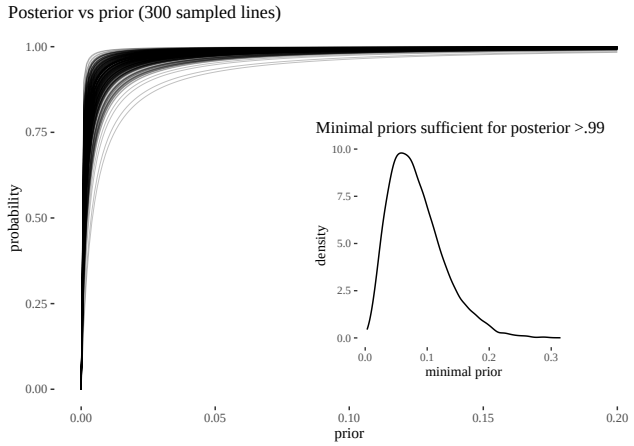

Figure 4: 300 lines illustrating the uncertainty about the dependence of the posterior on the prior given aleatory uncertainty about the evidence, with the distribution of the minimal priors required for the posterior to be above .99.

This, then, is the main claim of this chapter: whenever density estimates for the probabilities of interest are available (and they should be available for match evidence and many other items of scientific evidence if the reliability of a given type of evidence has been properly studied), those densities should

be reported for assessing the strength of the evidence. This approach avoids hiding actual aleatory uncertainties under the carpet. It also allows for a balanced assessment of the evidence, whereas using point estimates or intervals may exaggerate or underestimate the value of the evidence.

In what follows, we expand on this idea in different directions. Section 2 engages with the philosophical debate about precise and imprecise probabilism. We argue that both options are problematic and should be superseded by a higher-order approach to probability whenever possible. Section **??** revisits a recent discussion in the forensic science literature. A prominent view has it that trial experts, even when they use densities, should present only first-order probabilities. We disagree and show that reasons of accuracy maximization sometimes recommend relying on higher-order probabilities. Section **??** turns to some legal applications of higher-order probabilism. We focus on two topics: first, the role of higher-order probabilities and false positive rates in the evaluation of DNA evidence; second, how complex bodies of evidence can be represented by what we call higher-order Bayesian networks.

Before we dive in, one more remark: ost of the time, mathematically, we do not propose anything radically new—we just put together some of the items from the standard Bayesian toolkit. The novelty is rather in our arguing that that these tools are under-appreciated in the legal scholarship and should be properly used to incorporate second-order uncertainties in evidence evaluation and incorporation. Perhaps a minor exception is our explication of the notion of weight, but even here many related notions are available in information theory, and the novelty here is not technical, but rather in the argument that they also are under-appreciated in legal scholarship.

added this par to preemt Kadane's style pickiness

2 Three Probabilisms

The introduction outlined three probabilistic approaches that one might take for assessing the value of the evidence presented at trial. The first approach uses precise probabilities; the second uses intervals; the third uses distributions over probabilities. By relying on an example featuring two items of match evidence, we suggested that the third approach is preferable. This section buttresses this claim by providing principled, philosophical reasons in favor of the third approach.

The three approaches we considered correspond (roughly) to three ways in which probabilities can be deployed to model a rational agent's fallible and evidence-based beliefs about the world. The first approach, known in the philosophical literature as precise probabilism, posits that an agent's credal state is modeled by a single, precise probability measure. The second approach, known as imprecise probabilism, replaces precise probabilities by sets of probability measures. The third approach, what we call higher-order probabilism, relies on distributions over parameter values. There are good reasons to abandon precise probabilism and endorse higher-order probabilism. Imprecise probabilism is a step in the right direction, but also suffers from too many difficulties of its own.

2.1 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 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, but biases of coins can be thought of as random match probabilities in the forensic context.)

> **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 cannot 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. 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. But your precise probability assessment will again be .5.

These examples suggest that precise probabilism is not appropriately responsive to evidence. 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.[8]

2.2 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.[9]

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:

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

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 $\mathsf{P}(\mathsf{heads}|E)$ we need to calculate $\mathsf{P}(E|\mathsf{heads})\mathsf{P}(\mathsf{heads})$ and divide it by $\mathsf{P}(E) = \mathsf{P}(E|\mathsf{heads})\mathsf{P}(\mathsf{heads}) + \mathsf{P}(E) = \mathsf{P}(E|\neg\mathsf{heads})\mathsf{P}(\neg\mathsf{heads})$. The tricky part is obtaining the conditional probabilities $\mathsf{P}(E|\mathsf{heads})$ and $\mathsf{P}(E|\neg\mathsf{heads})$ in a principled manner without explicitly going second-order, estimating the parameter value and 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

[8] Precise probabilism suffers from other difficulties. For example, it has problems with formulating a sensible method of probabilistic opinion aggregation Stewart & Quintana (2018). A seemingly intuitive constraint is that if every member agrees that X and Y are probabilistically independent, the aggregated credence should respect this. But this is hard to achieve if we stick to PP (Dietrich & List, 2016). For instance, a *prima facie* obvious method of linear pooling does not respect this. Consider probabilistic measures p and q such that $p(X) = p(Y) = p(X|Y) = 1/3$ and $q(X) = q(Y) = q(X|Y) = 2/3$. On both measures, taken separately, X and Y are independent. Now take the average, $r = p/2 + q/2$. Then $r(X \cap Y) = 5/18 \neq r(X)r(Y) = 1/4$.

[9] 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; Kyburg Jr & Teng, 2001). 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. As we have already noted, intervals do not contain probabilistic information sufficient to guide reasoning with multiple items of evidence. So we focus on IP, which is the more promising approach.

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.[10]

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.[11] But no matter what we think about belief inertia, other problems plague imprecise probabilism. Two more problems are particularly pressing.

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 performs any valuable task.[12]

Second, besides descriptive inadequacy, an even deeper, foundational problem exists for imprecise probabilism. This problem arises when we attempt to measure the accuracy of a representor set of probability measures. Workable *scoring rules* exists 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.

[10] 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 $[0, 1]$ in each. 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.

[11] Elkin (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.

[12] 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.

2.3 Higher-order Probabilism

There is, however, a view in the neighborhood that fares better: a second-order perspective. In fact, some of the comments by the proponents of imprecise probabilism tend to go in this direction. For instance, Seamus 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 (Bradley, 2012, p. 157):

> ... 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. Joyce (2005), in a paper defending imprecise probabilism, in fact uses a density over chance hypotheses to account for the notion of evidential weight. 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.

The key 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.[13] 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

$$\int_0^1 x f(x)\, dx$$

as the precise, object-level credence in X, 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.[14] For example, consider again the scenario in which the agent knows that the bias of the coin is either .4 or .6 but the former is three times more likely. Representing the agent's credal state with the expectation $\mathsf{P}(X) = .75 \times .4 + .25 \times .6 = .45$ would be inadequate as it would fail to capture the agent's belief that the two biases are uneven.

The higher-order approach can easily model all the challenging scenarios we discussed so far in the manner illustrated in Figure 5. In particular, the scenario in which the two biases of the coin are not equally likely—which imprecise probabilism cannot model—can be easily modeled within high-order probabilism by assigning different probabilities to the two biases.

Besides its flexibility in modelling uncertainty, higher-order probabilism does not fall prey to belief inertia. Consider a situation in which you have no idea about the bias of a coin. So you start with a uniform density over $[0, 1]$ as your prior. By using binomial probabilities as likelihoods, observing any non-zero number of heads will exclude 0 and observing any non-zero number of tails will exclude 1 from the basis of the posterior. The posterior distribution will become more centered around the parameter estimate as the observations come in. Figure 6 shows—starting with a uniform prior distribution— how the posterior distribution changes after successive observations of heads, heads again, and then tails.[15]

A further advantage of high-order probabilism over imprecise probabilism is that the prospects for

[13] Bradley admits this much (Bradley, 2012, p. 90), 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.

[14] 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. Whether the expectation should be used in betting behavior is a separate problem. Here we focus on epistemic issues.

[15] More generally, learning about frequencies, assuming independence and constant probability for all the observations, is modeled the Bayes way. You start with some prior density p over the parameter values. If you start with complete lack of information, p should be uniform. Then, you observe the data D which is the number of successes s in a certain number of observations n. For each particular possible value θ of the parameter, the probability of D conditional on θ follows the

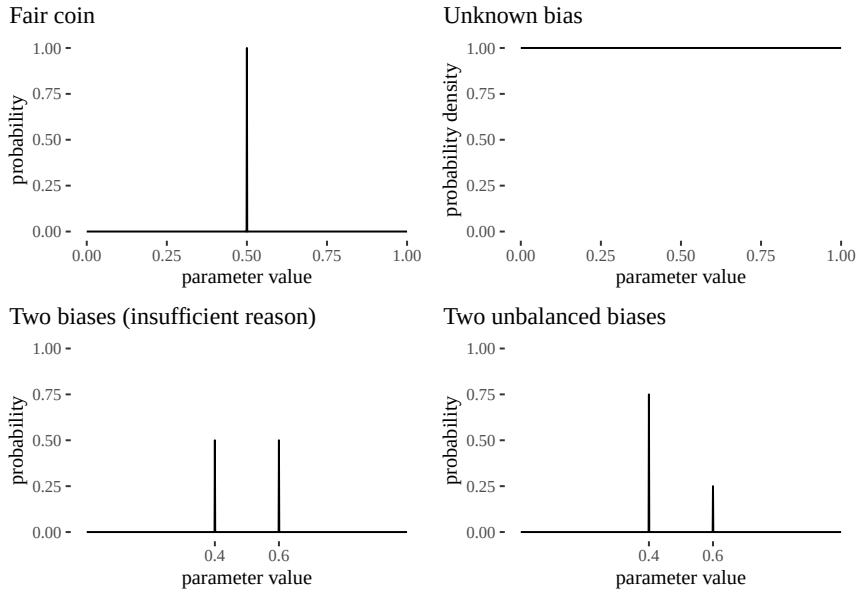

Figure 5: Examples of higher-order distributions for scenarios brought up in the literature.

accuracy-based arguments are not foreclosed. This is a significant shortcoming of imprecise probabilism, especially because such arguments exist for precise probabilism. One can show that there exist proper scoring rules for higher-order probabilism. These rules can then be used to formulate accuracy-based arguments. Another interesting feature of the framework is that the point made by Schoenfield against imprecise probabilism does not apply: there are cases in which accuracy considerations recommend an imprecise stance (that is, a multi-modal distribution) over a precise one (Urbaniak, 2022 manuscript).

All in all, higher-order probabilism outperforms both precise and imprecise probabilism, at the descriptive as well as the normative level. From a descriptive standpoint, higher-order probabilism can easily model a variety of scenarios that cannot be adequately modeled by the other versions of probabilism. From a normative standpoint, accuracy maximization may sometimes recommend that a rational agent represent their credal state with a distribution over probability values rather than a precise probability measure (more on this in the next section).

References

Bradley, S. (2012). *Scientific uncertainty and decision making* (PhD thesis). London School of Economics; Political Science (University of London).

Bradley, S. (2019). Imprecise Probabilities. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2019). https://plato.stanford.edu/archives/spr2019/entries/imprecise-probabilities/; Metaphysics Research Lab, Stanford University.

Campbell-Moore, C. (2020). *Accuracy and imprecise probabilities*.

Carr, J. R. (2020). Imprecise evidence without imprecise credences. *Philosophical Studies*, *177*(9), 2735–2758. https://doi.org/10.1007/s11098-019-01336-7

Deadman, H. A. (1984a). Fiber evidence and the wayne williams trial (conclusion). *FBI L. Enforcement Bull.*, *53*, 10–19.

binomial distribution. The probability of D is obtained by integration. That is:

$$\begin{aligned} p(\theta|D) &= \frac{p(D|\theta)p(\theta)}{p(D)} \\ &= \frac{\theta^s(1-\theta)^{(n-s)}p(\theta)}{\int(\theta')^s(1-\theta')^{(n-s)}p(\theta')\,d\theta'}. \end{aligned}$$

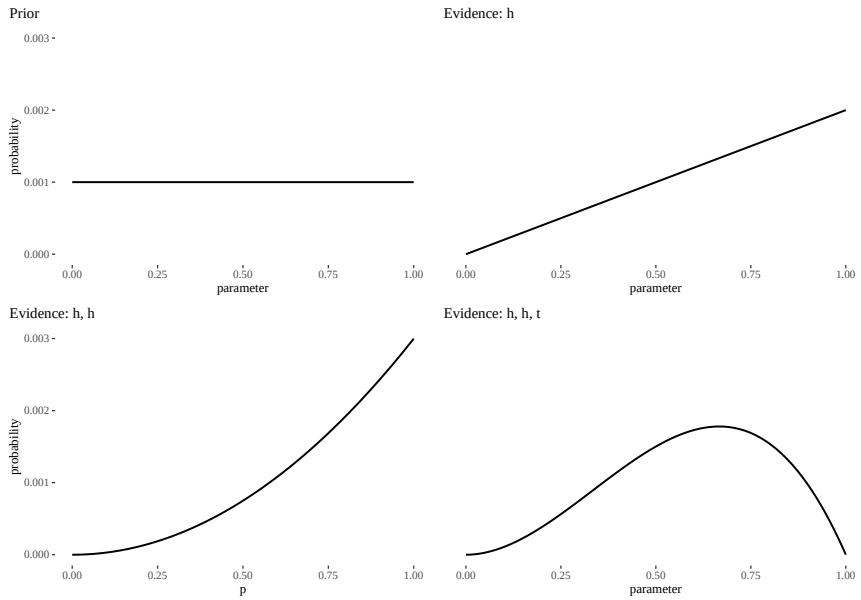

Figure 6: As observations of heads, heads and tails come in, extreme parameter values drop out of the picture and the posterior is shaped by the evidence.

Deadman, H. A. (1984b). Fiber evidence and the wayne williams trial (part i). *FBI L. Enforcement Bull.*, *53*, 12–20.

Dietrich, F., & List, C. (2016). Probabilistic opinion pooling. In A. Hajek & C. Hitchcock (Eds.), *Oxford handbook of philosophy and probability*. Oxford: Oxford University Press.

Elkin, L. (2017). *Imprecise probability in epistemology* (PhD thesis). Ludwig-Maximilians-Universität; Ludwig-Maximilians-Universität München.

Elkin, L., & Wheeler, G. (2018). Resolving peer disagreements through imprecise probabilities. *Noûs*, *52*(2), 260–278. https://doi.org/10.1111/nous.12143

Fraassen, B. C. V. (2006). Vague expectation value loss. *Philosophical Studies*, *127*(3), 483–491. https://doi.org/10.1007/s11098-004-7821-2

Gärdenfors, P., & Sahlin, N.-E. (1982). Unreliable probabilities, risk taking, and decision making. *Synthese*, *53*(3), 361–386. https://doi.org/10.1007/bf00486156

Joyce, J. M. (2005). How probabilities reflect evidence. *Philosophical Perspectives*, *19*(1), 153–178.

Kaplan, J. (1968). Decision theory and the fact-finding process. *Stanford Law Review*, *20*(6), 1065–1092.

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

Konek, J. (2013). *New foundations for imprecise bayesianism* (PhD thesis). University of Michigan.

Kyburg, H. E. (1961). *Probability and the logic of rational belief*. Wesleyan University Press.

Kyburg Jr, H. E., & Teng, C. M. (2001). *Uncertain inference*. Cambridge University Press.

Levi, I. (1974). On indeterminate probabilities. *The Journal of Philosophy*, *71*(13), 391. https://doi.org/10.2307/2025161

Levi, I. (1980). *The enterprise of knowledge: An essay on knowledge, credal probability, and chance*. MIT Press.

Mayo-Wilson, C., & Wheeler, G. (2016). Scoring imprecise credences: A mildly immodest proposal. *Philosophy and Phenomenological Research*, *92*(1), 55–78. https://doi.org/10.1111/phpr.12256

Rinard, S. (2013). Against radical credal imprecision. *Thought: A Journal of Philosophy*, *2*(1), 157–165. https://doi.org/10.1002/tht3.84

Schoenfield, M. (2017). The accuracy and rationality of imprecise credences. *Noûs*, *51*(4), 667–685. https://doi.org/10.1111/nous.12105

Seidenfeld, T., Schervish, M., & Kadane, J. (2012). Forecasting with imprecise probabilities. *International Journal of Approximate Reasoning*, *53*, 1248–1261. https://doi.org/10.1016/j.ijar.2012.06.018

Stewart, R. T., & Quintana, I. O. (2018). Learning and pooling, pooling and learning. *Erkenntnis*,

83(3), 1–21. https://doi.org/10.1007/s10670-017-9894-2

Sturgeon, S. (2008). Reason and the grain of belief. *Noûs*, *42*(1), 139–165. Retrieved from http://www.jstor.org/stable/25177157

Walley, P. (1991). *Statistical reasoning with imprecise probabilities*. Chapman; Hall London.