

Second-order Probabilism: Expressive Power and Accuracy

Rafal Urbaniak and Marcello Di Bello

2023-12-05

Table of contents

1	Introduction	2
2	Precise probabilism	2
3	Imprecise probabilism	3
4	Higher-order probabilism	6
5	Proper scoring for higher-order probabilism	9
6	Evidence aggregation	14
7	Computational and representational considerations	19
8	Discussion	22
	Appendix: propriety	24
	Appendix: the strict propriety of $\mathcal{J}_{D_{\text{KL}}}^2$	24
	References	27

DISCLAIMER: This is a draft of work in progress, please do not cite or distribute without permission.

1 Introduction

As rational agents, we form beliefs about a variety of propositions on the basis of the evidence available to us. But believing a proposition is not an all-or-nothing affair; it is a matter of degrees. We are uncertain, to a greater or lesser extent, about the truth of many propositions since the evidence we possess about them is often fallible. To represent this uncertainty, it is natural to use a probability measure that assigns to each proposition a value between 0 and 1 (also called a degree of belief or credence). This approach—known as *precise probabilism*—models an agent's state of uncertainty (or credal state) with a single probability measure: each proposition is assigned one probability value (a sharp degree of belief). The problem is that a sharp probability measure is not expressive enough to distinguish between intuitively different states of uncertainty rational agents may find themselves in (§2). To avoid this problem, a *set* of probability measures, rather than a single one, can be used to represent the uncertainty of a rational agent. This approach is known as *imprecise probabilism*. It outperforms precise probabilism in some respects, but also runs into problems of its own (§3).

To make progress, this paper argues that the uncertainty of a rational agent is to be represented neither by a single probability measure nor a set of measures. Rather, it is to be represented by a higher-order probability measure, more specifically, a probability distribution over parameter values interpreted as probabilities. Call this view *higher-order probabilism*. We show that higher-order probabilism addresses all the problems and philosophical puzzles that plague both precise and imprecise probabilism (§4 and §5).

Moreover, Bayesian probabilistic programming already provides a fairly reliable implementation framework of this approach.

add struc-
ture descrip-
tionThink about
including
synergy

2 Precise probabilism

Precise probabilism (PP) holds that a rational agent's uncertainty about a proposition is to be represented as a single, precise probability measure. Bayesian updating regulates how the prior probability measure should change in light of new evidence that the agent learns. The updating can be iterated multiple times for multiple pieces of evidence considered successively. This is an elegant and simple theory with many powerful applications. Unfortunately, representing our uncertainty about a proposition in terms of a single, precise probability measure 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 featuring coin tosses should make the point clear.

No evidence v. fair coin You are about to toss a coin, but have no evidence 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 scenarios 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 precise probabilism. After the 10 tosses, you again assess the probability to be .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. Representing an agent's uncertainty by a precise probability measure can fail to track what an agent has learned from new evidence. Precise probabilism assigns the same probability in situations in which one's evidence is quite different: when no evidence is available about a 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). In fact, analogous problems also arise for evidence that the coin is not fair. Suppose the rational agent starts with a weak belief that the coin is .6 biased towards heads. They can strengthen that belief by tossing the coin repeatedly and observing, say, 60 heads in 100 tosses. But this improvement in their evidence is not mirrored in the .6 probability they are supposed to assign to *heads*.¹

add reference about sweetening

These problems generalize beyond cases of coin tossing. It is one thing not to know much about whether a proposition is true, for example, whether an individual is guilty of a crime. It is another thing to have strong evidence that favors a hypothesis and equally strong evidence that favors its negation, for example, strong evidence favoring the guilt hypothesis and equally strong evidence favoring the hypothesis of innocence. Despite this difference, precise probabilism would recommend that a probability of .5 be assigned to both hypotheses in either case. Here, too, precise probabilism fails to be appropriately responsive to the evidence. In addition, evidence can accumulate in a way that does not require changing our initial probability assignments. Suppose that, at first, one has stronger evidence overall for *A* than for *B*. So the probability assigned to *A* should be greater than the probability assigned to *B*. Next, the agent acquires new evidence. The total quantity of evidence has therefore increased, but this larger body of evidence overall still favors *A* over *B*. No change in the probabilities seems required. Still, something has changed about the agent's state of uncertainty towards *A* and *B*: the quantity of evidence on which the agent can make their assessment whether *A* is more probable than *B* has become larger. And yet, this change in the quantity of overall evidence is not reflected in the precise probabilities assigned to the propositions *A* and *B*.²

3 Imprecise probabilism

What if we give up the assumption that probability assignments should be precise? Imprecise probabilism (IP) holds that a rational agent's credal stance towards a hypothesis is to be represented by 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

¹Here is another problem for precise probabilism. Imagine a rational agent who does not know the bias of the coin. For precise probabilism, this state of uncertainty should be represented by a .5 probability assignment to the *heads*. Next, the agent learns that the bias towards heads, whatever the bias is, has been slightly increased, say by .001. The addition of this new information is called *sweetening* in the philosophical literature. This sweetening should now make the agent bet on heads: if the probability of *heads* was initially .5, it must be ever so slightly above .5 after sweetening. But, intuitively, the new information should leave the agent equally undecided about betting on heads or tails. After sweetening, the agent still does not know much about the actual bias of the coin.

²Following Keynes, the distinction here is between the *balance* of the evidence—whether the evidence available tips in favor of a proposition or another—and the *weight* of the evidence—the overall quantity of evidence. **ADD REFERENCE TO KEYNES**

are compatible with the evidence (more on this point later).³ It is easy to see that modeling an agent's credal state by sets of probability measures avoids some of the shortcomings of precise probabilism. For instance, if an agent knows that the coin is fair, their credal state would be represented by the singleton set $\{P\}$, where P is a probability measure that 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.

So far so good. But, just as precise probabilism fails to be appropriately evidence-responsive in certain scenarios, imprecise probabilism runs in similar difficulties in other scenarios.

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. Since the probability measures in the set are all compatible with the agent's evidence, no probability measure can be assigned a greater (higher-order) probability than any other.⁴

These examples show that imprecise probabilism is not expressive enough to model the scenario of uneven bias. Defenders of imprecise probabilism could concede this point but prefer their account for reasons of simplicity. They could also point out that imprecise probabilism models scenarios that precise probabilism cannot model, for example, a state of complete lack of evidence. In this respect, imprecise probabilism outperforms precise probabilism in expressive power, but also retains theoretical simplicity. Unfortunately, it is questionable whether imprecise probabilism actually outperforms precise probabilism *all things considered*. As we will now see, imprecise probabilism suffers from a number of shortcomings that do not affect precise probabilism.

The first problem has not received extensive discussion in the literature, but it is fundamental. Recall that, for imprecise probabilism, an agent's state of uncertainty is represented by those probability measures that are *compatible* with the agent's evidence. The question is, how should the notion of compatibility be understood here? Perhaps we can think of compat-

³For 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.

⁴Other 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 $P(M) = r$. But then, it also contains one with $P(M) = 1$. This means that it is not the case that for any probability measure P in your representor, $P(G) > P(M)$, that is, it is not the case that RA is more confident of G than of M . This is highly counter-intuitive.

ibility as the fact that the agent’s evidence is consistent with the probability measure in question. But mere consistency wouldn’t get the agent very far in excluding probability measures, as too many probability measures are strictly speaking still consistent with most observations and data. Admittedly, there will be clear-cut cases: if you see the outcome of a coin toss to be heads, you reject the measure with $P(H) = 0$, and similarly for tails. Another class of cases might arise while randomly drawing objects from a finite set where the true frequencies or objective chances are already known, because the finite set has been inspected. But such clear-cut cases aside, what else? In the end, evidence will often be consistent with almost any probability measure.⁵

A second, well-known problem for imprecise probabilism is belief inertia. Precise probabilism offers an elegant model of learning from evidence: Bayesian updating. Imprecise probabilism, at least *prima facie*, offers an equally elegant model of learning from evidence, richer and more nuanced. It is a natural extension of the classical Bayesian approach that uses precise probabilities. 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:

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

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 measures 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 narrower.

Unfortunately, this narrowing of the range of values becomes impossible whenever the starting point is complete lack of knowledge, as imprecise probabilism runs into the problem of belief inertia (Levi, 1980). This problem arises in situations in which no amount of evidence could lead the agent to change their belief state, according to a given modeling strategy. Consider a situation in which 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. For any sequence of outcomes that you can obtain and any probability value in $[0, 1]$, there will exist a probability measure (conditional on the outcomes) that assigns that probability to *heads*. 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.⁶

⁵Probability measures can be inconsistent with evidential constraints that agents believe to be true. Mathematically, non-trivial evidential constraints are easy to model (Bradley, 2012). They can take the form, for example, of the *evidence of chances* $\{P(X) = x\}$ or $P(X) \in [x, y]$, or *structural constraints* such as “ X and Y are independent” or “ X is more likely than Y .” These constraints are something that an agent can come to accept outright, but only if offered such information by an expert whom the agent completely defers to. But, besides these idealized cases, it is unclear how an agent could come to accept such structural constraints upon observation. There will usually be some degree of uncertainty about the acceptability of these constraints.

⁶Here’s another example of inertia, coming 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). Suppose your initial credence about these two hypothesis is complete uncertainty with interval. Next, suppose you learn that a marble drawn at random from the urn is green (E). After using this evidence to condition each probability measure in your representor (which initially contains all possible probability measures over the relevant space) on this evidence, you end up with the same spread of values for H_1 that you had before learning E . This holds no matter how many marbles are sampled from the urn and found to be green. This is counterintuitive: if you continue drawing green marbles, even if you started with complete uncertainty, you should become more inclined towards the hypothesis that all marbles are green.

Some downplay the problem of belief inertia. After all, if you started with knowing truly nothing, then it is right to conclude that you will never learn anything. Joyce (2010) writes:

You cannot learn anything in cases of pronounced ignorance simply because a prerequisite for learning is to have prior views about how potential data should alter your beliefs (p. 291)

The upshot is that vacuous priors should not be used and that imprecise probabilism gives the right results when the priors are non-vacuous. Another strategy is to say that, in a state of complete ignorance, a special updating rule should be deployed.⁷

Finally, imprecise probabilism faces a third, deeper problem that does not arise for precise probabilism. As it turns out, it is impossible to define proper scoring rules for measuring the accuracy of a representor set of probability measures. Workable *scoring rules* exist for measuring the accuracy of a single, precise probability measure, such as the Brier score. These rules measure the distance between a rational agent's probability measure (also called credence function) and the actual value. A requirement of scoring rules is that they be *proper*: any rational agent will expect their own probability measure to be more accurate than any other. After all, if an agent thought a different probability measure 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 measure follows the axioms of probability theory, no other non-probabilistic measure is going to be more accurate than yours whatever the facts are. Can the same be done for imprecise probabilism? It cannot. 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.

add reference. Joyce, J. M. (2010). A Defence of Imprecise Credences in Inference and Decision Making. **Philosophical Perspectives** 24, pp. 281–323.

Rafal to develop a response to this point.

4 Higher-order probabilism

Let us take stock. Imprecise probabilism is more expressive than precise probabilism. It can model the difference between a state in which there is no evidence about a proposition (or its negation) and a state in which the evidence for and against a proposition is in equipoise. But imprecise probabilism has its own expressive limitations: it cannot model the case of uneven bias. In addition, imprecise probabilism faces difficulties that do not affect precise probabilism: the notion of compatibility between a probability measure and the evidence is too permissive; belief inertia makes it impossible for a rational agent to learn via Bayesian updating; and no proper scoring rules exist for imprecise probabilism. In this section, we show that higher-order probabilism overcomes the expressive limitations of imprecise probabilism without falling prey to any such difficulties.

Proponents of imprecise probabilism already hinted to the need of relying on higher order probabilities. For instance, Bradley compares the measures in a representor to committee members, each voting on a particular issue, say the true chance or bias of a coin. As they acquire more evidence, the committee members will often converge on a chance hypothesis.

...the committee members are “bunching up”. Whatever measure you put over the

⁷Elkin (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 from a principled account of why one should use a separate special update rule when starting with complete ignorance.

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. (Bradley, 2012, p. 157)

But such bunching up cannot be modeled by imprecise probabilism alone: a probability distribution over chance hypotheses is needed.⁸ 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 and imprecise probabilism. We now set out to do precisely that.

The central idea of higher-order probabilism is this: a rational agent’s uncertainty is not single-dimensional and thus cannot be mapped onto a one-dimensional scale like the real line. Uncertainty is best modeled by the shape of a probability distribution over multiple probability measures. Stated more formally, a rational agent’s state of uncertainty (or credal stance) towards a proposition X is not represented by a single probability value $P(X)$ between 0 and 1, but by a probability (density) distribution $f_{P(X)}$, where the first-order probability measure $P(X)$ is treated as a random variable. Crucially, this representation is completely general. Unlike the examples used so far, the proposition X is not restricted to chance hypotheses or the bias of a coin. The probability distribution $f_{P(X)}$ assigns a second-order probability to each of the first-order probabilities $P(X)$.

How should these second-order probabilities be understood? It is helpful to think of higher-order probabilism as a generalization of imprecise probabilism. Imprecisers already admit that some probability measures are compatible and others incompatible with the agent’s evidence at some point. Compatibility is a coarse notion; it is an all-or-nothing affair. But, as seen earlier, evidence can hardly exclude a probability measure in a definitive manner except in clear-cut cases. Just as it is often a matter of degrees whether evidence supports a proposition, the notion of compatibility between evidence and probability measures can itself be a matter of degrees. On this picture, the evidence justifies different values of first-order probability to various degrees. So, second-order probabilities express the extent to which the first-order probabilities are supported by the evidence.

This higher-order approach at the technical level is by no means novel. Bayesian probabilistic programming languages embrace the well-known idea that parameters can be stacked and depend on each other (Bingham et al., 2021). But, while the technical machinery has been around for a while, it has not been deployed by philosophers to model a rational agent’s uncertainty or credal state. Because of its greater expressive power, higher-order probabilism can represent uncertainty in a more fine-grained manner, as illustrated in Figure 1. In particular, the uneven coin 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.

An agent’s uncertainty could—perhaps, should—sometimes be represented by a single probability value. Higher-order probabilism does not prohibit that. For example, there may well be cases in which an agent’s uncertainty is aptly represented by the expectation.⁹ But this need not always be the case. If the probability distribution is not sufficiently concentrated around a single value, a one-point summary will fail to do justice to the nuances of the agent’s

Rafal to correct technical mistakes in this paragraph (e.g. concept of distribution over probability measures seems incorrect for our purposes)

⁸In a similar vein, Joyce (2005), in a paper defending imprecise probabilism, explicates the notion of weight of evidence using a probability distribution over chance hypotheses. Oddly, representor sets play no central role in Joyce’s account of the weight of evidence.

⁹The expectation is usually defined as $\int_0^1 x f(x) dx$. In the context of our approach here, x is the first-order probability of a given proposition, and f is the density representing the agent’s uncertainty about x .

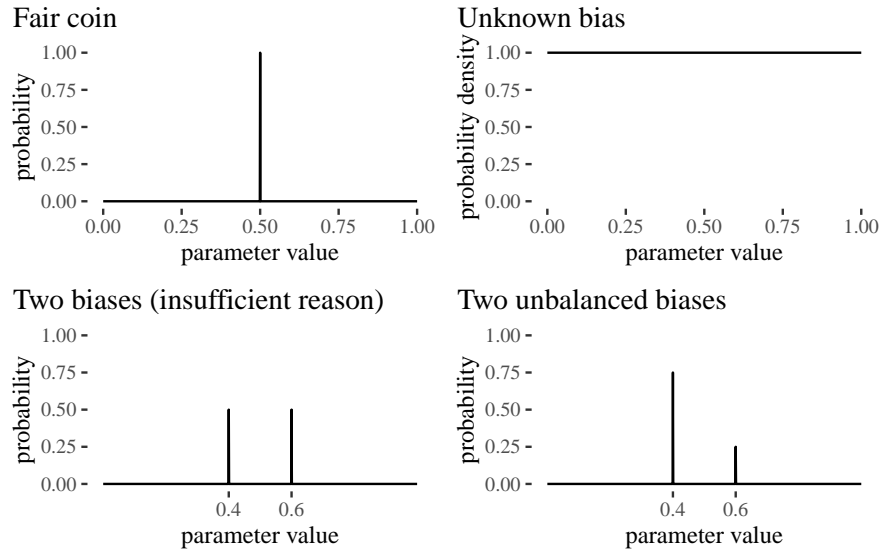


Figure 1: Examples of higher-order distributions for a few scenarios problematic for both precise and imprecise probabilism.

credal state.¹⁰ 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 $P(X) = .75 \times .4 + .25 \times .6 = .45$ would fail to capture the agent’s different epistemic attitudes towards the two biases. The agent believes the two biases have different probabilities, but is also certain the bias is *not* .45.

Besides its greater expressive power in modelling uncertainty, higher-order probabilism does not fall prey to belief inertia or the impossibility of proper scoring rules. Consider a situation in which you have no idea about the bias of a coin. You start with a uniform distribution over $[0, 1]$ as your prior. 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 as the observations come in. This result is a straightforward application of Bayesian updating. Instead of plugging sharp probability values into the formula for Bayes’s theorem, the factors to be multiplied in the theorem will be probability densities (or ratios of densities as needed). Figure 2 illustrates—starting with a uniform prior distribution—how the posterior (beta) distribution changes after successive observations of heads, heads again, and then tails.¹¹

The impossibility of proper scoring rules was another weakness of imprecise probabilism. This is a significant shortcoming, especially because proper scores do exist for precise proba-

¹⁰This 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.

¹¹Assuming independence and constant probability for all the observations, learning is modeled the Bayesian 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 binomial distribution. The probability of D is obtained by integration. That is:

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

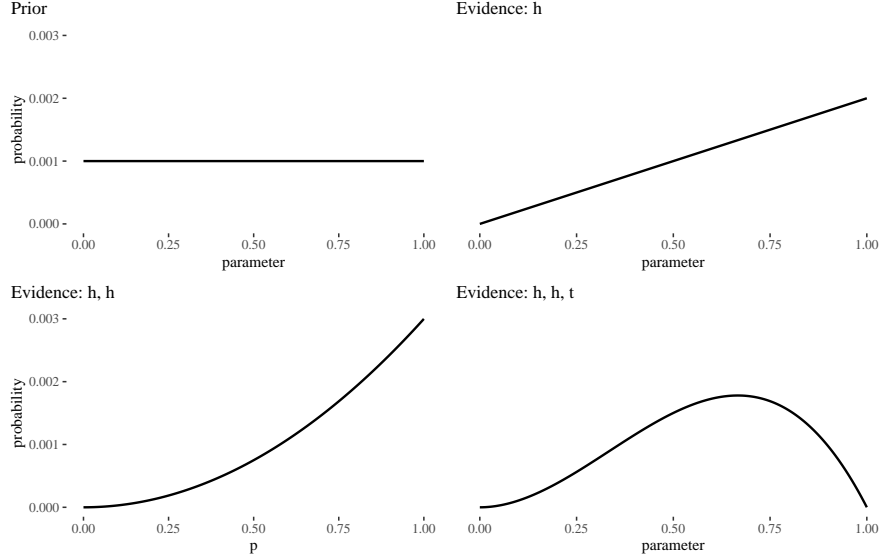


Figure 2: As observations of heads, heads and tails come in, extreme parameter values drop out of the picture and the posterior is shaped by the evidence.

bilism. Fortunately, one can show that there exist proper scoring rules for higher-order probabilism. These rules can then be used to formulate accuracy-based arguments. In addition, recall the point made by Schoenfield (2017): an accuracy measure will not usually recommend an imprecise stance. This argument fails against imprecise probabilism: there are cases in which accuracy considerations recommend an imprecise stance (that is, a multi-modal distribution) over a precise one. We will defend these claims in the next section. It relies more heavily on a formal apparatus, and it can be skipped without losing track of the main line of the argument.

5 Proper scoring for higher-order probabilism

As already noted, one challenge for the imprecisers is providing a proper scoring rule that would be a counterpart of, say, the Brier score for the precise case. Imprecise probabilism cannot meet this challenge, but we show that higher-order probabilism can.

The task is to define a score for comparing the accuracy of probability distributions. As a first pass, we can help ourselves to existing work on the accuracy of continuous probability distributions (Hersbach (2000), Pettigrew (2012), Gneiting & Raftery (2007)). The continuous ranked probability score (CRPS) that assesses how well a probability distribution aligns with a chance hypothesis that either puts all weight on 0 or 1 can be of service here. The (CRPS) of a distribution p with respect to a possible world w is defined as follows:

$$I(p, w) = \int_{-\infty}^{\infty} |P(x) - \mathbf{1}(x \geq V(w))|^2 dx$$

where P is the cumulative probability corresponding to the probability distribution p , and

$$\mathbf{1}(x \geq V(w)) = \begin{cases} 1 & \text{if } x \geq V(w) \\ 0 & \text{o/w.} \end{cases}$$

The CRPS score measures the distance relative to an epistemically omniscient chance hypoth-

esis, which either puts full weight on 0, if a given proposition is false, or on 1, otherwise.¹² Measuring the distance between the cumulative distributions helps capture the overall disparity between the predicted and actual probabilities. The CRPS generalizes the Brier score as the distance between the true value and the assigned probability to the case of continuous probability distributions. We will start by reflecting on this approach, but we will soon see that it runs into difficulties.

For computational ease, we will be using a grid approximation of the densities, as in practice we are unable to work with infinite precision anyway (note for instance that there are no readily computable solutions to the integral used in the definition of CRPS, although it can sometimes be evaluated in closed form) (Gneiting & Raftery, 2007, p. 366) In parallel, we will be also using the Kullback-Leibler divergence:

$$D_{\text{KL}}(p \parallel q) = \sum_x p(x) \log \left(\frac{p(x)}{q(x)} \right)$$

which is a standard information-theoretic measure of divergence of q from p from the perspective of p .¹³

To fix ideas, consider a variation of a scenario by Schoenfeld (2017). A rational agent is invited to engage in a bet by an opponent who has two coins. The behavior of the coins, instead of being described with single point estimates, is described with parameters that reflect the shape of the distribution of n tosses. One of these coins has a normal distribution of *Heads* centered at .3, while the other is centered at .5. Both coins have a standard deviation of .05. The opponent randomly selects one of these coins and flips it. The rational agent knows all the details of this set-up.

What credal state should the rational agent form in response? Consider three options, but there could be more: first, a *faithful bimodal* distribution centered at .3 and .5; second, a *unimodal* distribution centered at .4; third, a *wide bimodal* distribution centered at .2 and .6. The three options are depicted in Figure 3. All of them have expected values at .4 rounded to four digits.

¹²The CRPS score is a particular case of the more general Cramer-Von-Mises measure of distance between densities, defined in terms of the area under the squared euclidean distances between the corresponding cumulative density functions:

$$\mathcal{C}(p, q) = \int_0^1 |P(x) - Q(x)|^2 dx$$

¹³In the continuous case we'd need to use differential KL divergence.

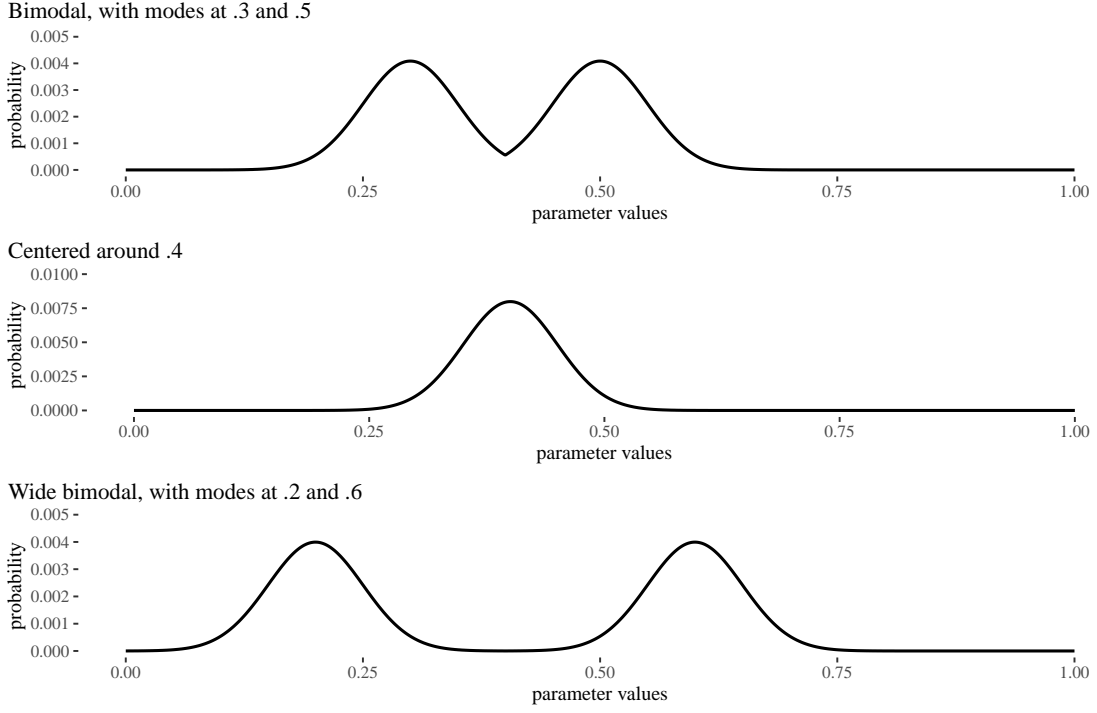


Figure 3: Three distributions in a vague EMS scenario. The distributions are built from normal distributions with standard deviation .5, the bimodal ones are “glued” in the middle.

Denote these distributions as b, c, w correspondingly. Now consider what happens if we think of expected inaccuracy of these distributions with possible true outcomes, conceptualized as either heads (H) or tails (T). Whatever our inaccuracy measure, we will have six inaccuracy scores of the form $I(\text{distribution}, \text{outcome})$, where outcome is one of two omniscient distributions that give all weight to either heads or tails. In the calculation of expected values, these will need to be multiplied by the probability of H and the probability of T . However, none of these distributions assigns such a probability, so to obtain such point values we in fact already need to take expected values of the form $\mathbb{E}_{\text{distribution}}(H) = \sum(x * \text{distribution}(x))$, where x are the values on the discretized grid (as in our example these expectations are pretty much the same, we can simply take $\text{distribution}(H)$ to be .4):

$$\mathbb{E}I_{\text{binary}}(p, q) = I(p, \text{heads})\mathbb{E}q(\text{heads}) + I(p, \text{tails})\mathbb{E}q(\text{tails}).$$

Table 1 displays the accuracy scores assigned by CRPS assuming the true outcome is *heads* (CRPS1) and assuming the true outcome is *tails* (CRPS0), and analogously for the KL divergence. Then, we obtain the expected inaccuracies by averaging the two scores by using the point probability of heads set to .4. These result in ExpCRPS and ExpKL respectively, note that since the probability of heads is the same on all the distributions, those are expected values from the perspective of each of the measures; changing the perspective in this example doesn’t change the expected inaccuracy.

distribution	CRPS1	CRPS0	KLD1	KLD0	ExpCRPS	ExpKLD
bimodal	534.7305	334.9305	80.06971	33.90347	414.8505	52.36997
centered	571.2192	371.4192	110.84220	53.13440	451.3392	76.21752
wide bimodal	485.4052	285.6177	54.13433	19.50965	365.5340	33.35974

Table 1: CPRS and KLD inaccuracies of the three distributions to the TRUE and FALSE omniscient functions, with expected inaccuracies.

As it turns out, the expected CRPS score recommends the wide bimodal distribution as the most accurate (or least inaccurate). The KL divergence from the omniscient measure makes the same recommendation.¹⁴ This is counterintuitive because the faithful bimodal seems the one most appropriately evidence-responsive. The unimodal distribution, while centering on the expected value, gets the chances wrong, and the wide bimodal has its guesses too close to the truth values and too far from the known chances. But there is a further problem. While the wide bimodal distribution expects itself to be the least inaccurate, the other distributions also expect the wide bimodal to be the least inaccurate. This indicates that in this setting the CRPS score and the KL divergence are not proper scores, as they allow for cases of some distributions recommending other distributions as less accurate, whatever the true state of the world turns out to be.

What are we to make of this result? Note that the three distributions share the same expected value .4. The latter is then used in the calculations of the expected inaccuracies for both CRPS and KL. The approach consists of (i) calculating two distances from the two extreme omniscient measures and (ii) averaging by plugging in the expected value. This approach, however, runs against the spirit of our enterprise. If expected values are often not good representations of a rational agent’s uncertainty, it should not be surprising that relying on them fails to deliver plausible expected accuracy scores. By reducing each of the distributions’ stance towards heads to a single point value .4, we’ve effectively washed out key information. So the question is, how can we adequately account for the complexity of a rational agent’s credal state in formulating a proper accuracy score?

Here is our proposal. Rather than measuring inaccuracy in relation to “true states of the world” conceptualized as two omniscient credences that peak at either 0 or 1 and then averaging using expected values, we should instead utilize a set of n potential true probability hypotheses (ideally, going continuous, but we’re working with a discrete grid of $n = 1000$ possible coin baisses in this paper). We then compute all the inaccuracies with respect to each of these n values represented by “omniscient” distributions (or true chance hypotheses) and determine the expected inaccuracy scores using the entire distributions rather than relying solely on their expected values.

$$\mathbb{E}I_{discretized}(p, q) = \sum I(p, x)q(x),$$

where $I(p, x)$ denotes the inaccuracy score of distribution p from the omniscient distribution assigning all weight to x , for x belonging to the discretized grid.

For the three distributions under consideration, the accuracy scores calculated using CRPS and KL divergence with respect to omniscient distributions corresponding to various values of x are given by Figure 4. The expected inaccuracies of the distributions from their perspective are given by Table 2. The results now match our common sense: each distribution recommends itself. So, once we pay attention to the whole range of possibilities, the CRPS score and KL divergence are now proper scores.

¹⁴This indicates that the choice of the evaluation metric is not the cause of the recommendation.

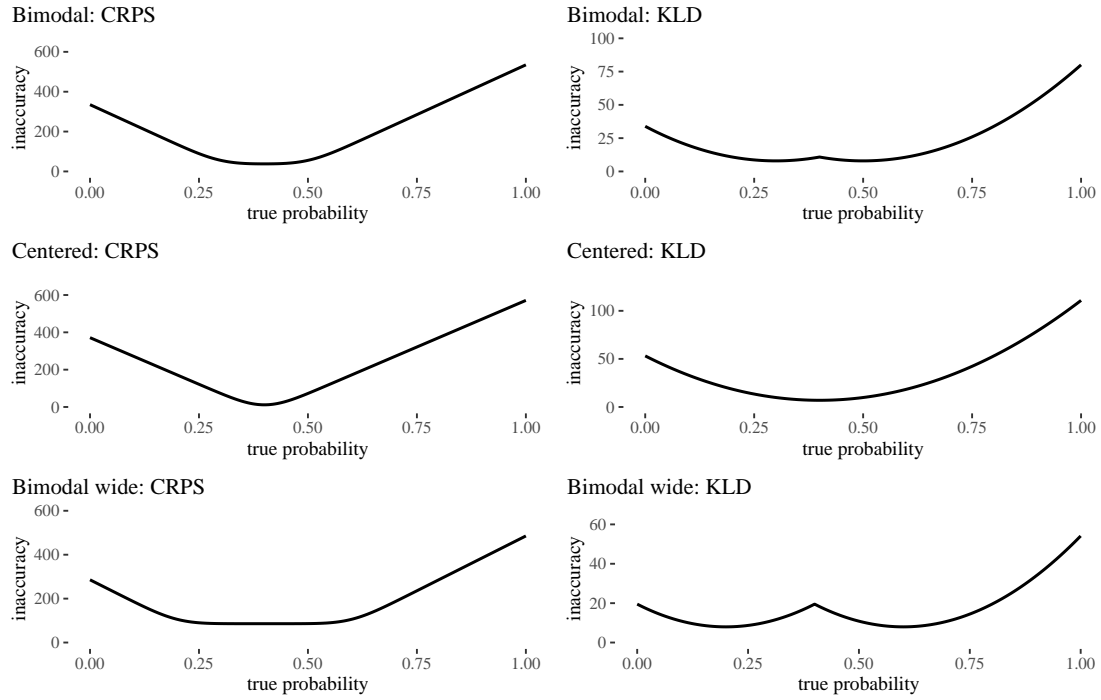


Figure 4: CLPSR and KL divergence based inaccuracies vs (omniscient functions corresponding to) n true probability hypotheses for the three distributions discussed in this section.

	CRPS			KLD		
	bimodal	centered	wide bimodal	bimodal	centered	wide bimodal
bimodal	64.670	78.145	88.380	8.577	10.655	11.336
centered	41.657	28.181	85.911	9.239	7.690	15.627
wide bimodal	137.699	171.719	113.989	11.541	19.231	8.689

Table 2: Expected inaccuracies of the three distributions from their own perspectives. Each row corresponds to a perspective.

One important difference transpires between using CRPS and KLD. Notice how for chance hypotheses between the actual peaks the inaccuracy remains flat. This seems to be an artifice of choosing a squared distance metric. If instead we go with a more principled, information-theory-inspired KL divergence, inaccuracy in fact jumps a bit for values in between the peaks for the bimodal distributions, which seems intuitive and desirable. This seems to be a reason to prefer a KL-based accuracy measure.

One question remains: how does the framework capture the idea that it is the bimodal distribution that seems more adequate than the others? One way to interpret that is by looking at inaccuracy concerning chance hypotheses given by the testimonial evidence. In this case, these are H_3 , where the true chance is 0.3, and H_5 , where the true chance is 0.5. You can find the specific inaccuracies for them in Table 3. To make sure that this favorable outcome isn't due to not using pointed credences, we can redo the calculations using the pointed version. In the pointed version, all the focus is on 0.4, or the weight is evenly divided between 0.3 and 0.5, or between 0.2 and 0.6. As anticipated, when we consider inaccuracy, both of these setups recommend the bimodal version (Table 4).

	CRPS		KLD	
	H3	H5	H3	H5
bimodal	55.475	55.378	7.935	7.935
centered	72.281	72.090	9.836	9.825
wide bimodal	86.230	86.223	10.871	10.882

Table 3: CRPS and KLD inaccuracies of the three distributions with respect to the two hypotheses. On both inaccuracy measures the bimodal distribution dominates the other two.

	CRPS		KLD	
	H3	H5	H3	H5
pointed bimodal	49.75	49.75	1.00	1.00
pointed centered	100.00	100.00	16.61	16.61
pointed wide bimodal	99.75	99.75	16.61	16.61

Table 4: CRPS and KLD inaccuracies of the three-pointed distributions with respect to the two hypotheses.

The discussion so far, while based on an example, may raise questions about the strict propriety of the KLD as an inaccuracy measure. To address this concern, a proof is provided in the appendix. In a nutshell, the argument demonstrates that for a second-order discretized probability mass p over a parameter space $[0, 1]$, with the actual probability denoted as θ , the Kullback-Leibler divergence of p from the indicator distribution of θ (which assigns 1 to θ and 0 to all other parameter values in the parameter space) is expressed as $\mathcal{J}_{D_{KL}}^2$.¹⁵ This serves as a demonstration of the strict propriety of the inaccuracy measure: each p anticipates itself to be the least inaccurate distribution.¹⁶

6 Evidence aggregation

The upshot so far is this. 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 soon).¹⁷ In this final section, we present a new challenge for imprecise probabilism, and we

¹⁵The argument generalizes to parameter spaces that correspond to probabilities of multiple propositions which are Cartesian products of parameter spaces explicitly used in the argument in this section.

¹⁶The argument has four key moves:

1. the inaccuracy of p w.r.t. to parameter θ is just $-\log_2 p(\theta)$,
2. the expected inaccuracy of p from the perspective of p is the entropy of p , $H(p)$,
3. the inaccuracy of q from the perspective of p is the cross-entropy $H(p, q)$,
4. and it is an established result that cross-entropy is strictly larger than entropy as soon as $p \neq q$.

¹⁷Having read van Fraassen’s “Laws and Symmetry”, you might also worry that going higher order somehow leads to a contradiction; we will address this concern later on.

add ref to vFraasen in fn; perhaps extend the discussion a bit

where do we show that accuracy recommends a distribution over probability measures? This bit seems missing.

show that higher-order probabilism can meet this challenge.

Rational agents are often tasked with aggregating pieces of evidence and assessing their value relative to a hypothesis. For two items of evidence, E_a and E_b , the posterior odds are obtained by multiplying the likelihood ratio by the prior odds while filling in the required probabilities:

$$\frac{P(H|E_a \wedge E_b)}{P(\neg H|E_a \wedge E_b)} = \frac{P(E_a \wedge E_b|H)}{P(E_a \wedge E_b|\neg H)} \times \frac{P(H)}{P(\neg H)}$$

This is a straightforward application of Bayes's theorem in odds form (often easier to use while carrying out calculations.) In the simplest case, the pieces of information are independent lines of evidence both relevant for hypothesis H . Stated formally, E_a and E_b are probabilistic independent conditional on hypothesis H . Think, for example, at two diagnostic tests performed by two independent laboratories or two independent witnesses in a trial testifying about the same issue. The posterior odds results from multiplying the likelihood ratios associated with each piece of evidence by the prior odds:

$$\frac{P(H|E_a \wedge E_b)}{P(\neg H|E_a \wedge E_b)} = \frac{P(E_a|H)}{P(E_a|\neg H)} \times \frac{P(E_b|H)}{P(E_b|\neg H)} \times \frac{P(H)}{P(\neg H)}$$

In this simple case, precise probabilism seems well equipped to handle evidence aggregation. (We will look at more complex scenarios later.)

Unfortunately, precise probabilities are not always available. Even when they are, they are not necessarily the best way to capture the uncertainty of the situation. An example can illustrate this point. Consider a murder case in which the police recover two items of trace evidence, both against the defendant. First, hair found at the crime scene matches the defendant's hair; call this evidence hair match. Second, the fur of the defendant's dog matches the fur found in a carpet wrapped around one of the bodies; call this evidence fur match.¹⁸ The two matches favor the hypothesis that the defendant (and the defendant's dog) must be the source of the crime traces; call this hypothesis source). If the two matches are independent lines of evidence (as they seem to be), their likelihood ratios can be multiplied:¹⁹

$$\frac{P(\text{dog} \wedge \text{hair}|\text{source})}{P(\text{dog} \wedge \text{hair}|\neg \text{source})} = \frac{P(\text{dog}|\text{source})}{P(\text{dog}|\neg \text{source})} \times \frac{P(\text{hair}|\text{source})}{P(\text{hair}|\neg \text{source})}$$

In assigning precise probabilities, the numerators can simply be equated to one. If the defendant is a contributor, the laboratory will declare a match for sure. To fill in the denominators, an expert will provide the relevant random match probabilities, for example, by counting of many matches are found in a representative sample of the population. Suppose the matching hair type occurs 0.0253 times in a reference database, and the matching dog fur type occurs 0.0256 times in a reference database.²⁰ These frequencies can be used to fill in the random match probabilities. So, putting everything together:

$$\frac{P(\text{dog}|\text{source})}{P(\text{dog}|\neg \text{source})} \times \frac{P(\text{hair}|\text{source})}{P(\text{hair}|\neg \text{source})} = \frac{1}{0.0252613} \times \frac{1}{0.025641} = \frac{1}{6.4772626 \times 10^{-4}}$$

¹⁸The hair evidence and the dog fur evidence are stylized after two items of evidence in the notorious 1981 Wayne Williams case (Deadman, 1984b, 1984a).

¹⁹It is possible for A and B to be independent conditional on C , but not conditional on $\neg C$. Here, we require both independencies to hold.

²⁰Probabilities have been slightly but not unrealistically modified to be closer to each other in order to make a conceptual point: the same first-order probabilities, even when they sound precise, may come with different degrees of second-order uncertainty (more on this soon). The original probabilities were 1/100 for the dog fur, and 29/1148 for Wayne Williams' hair.

This is a large number. The combined match evidence is strongly incriminating. It would be a huge coincidence if both matches were due to mere chance.

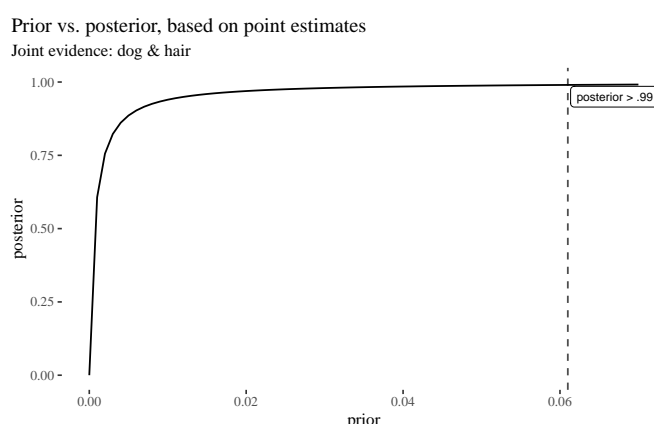


Figure 5: Impact of dog fur and human hair evidence on the prior, point estimates.

So far so good. But precisely probabilism quickly runs into trouble here. First, in order to find out the posterior odds, the likelihood ratios must be multiplied by the prior odds $\frac{P(\text{source})}{P(\neg\text{source})}$. What precise value should they take? This question is hard to answer definitively. To circumvent this difficulty, Figure 5 shows how the items of match evidence, combined, change the probability of the source hypothesis given a range of possible priors. This strategy—known as *sensitivity analysis*—shows that the posterior of .99 is reached as soon as the prior is higher than 0.061.²¹

Sensitivity analysis moves away from precise probabilism. It gives up on the prospect of assigning precise numbers to the prior probabilities. The second problem with precise probabilism here is that the match probabilities $P(\text{dog}|\neg\text{source})$ and $P(\text{fur}|\neg\text{source})$ are also subject to uncertainty. They are assessed by looking at sample data. So why pick exactly the numbers 0.0253 and 0.0256 for the match probabilities? Could they be different?

Imprecise probabilism might seem well equipped to model the uncertainty about match probabilities. But this is not so. In imprecise probabilism, the probability measures in the representor set are those compatible with the evidence, data or observations. Virtually any precise random match probability is compatible with any sample data—with any number of matches found in a reference database. Think by analogy to coin tossing: even a coin that has a .99 bias toward tails could come up heads on every toss. This series of outcomes is unlikely, but still possible. Similarly, even a hair type that has a match probability extremely small could be found several times in a sample population. So, from the perspective of imprecise probabilism, it is not clear how to proceed forward if one takes seriously the binary notion of compatibility. Imprecise probabilism is so permissive that any match probability will count as compatible with the data.

Another option is to identify reasonable ranges of probabilities. What are the worst-case and best-case scenarios? Suppose the reasonable ranges of the match probabilities are (.015, .037) (.002, .103), for hair and fur evidence respectively.²² It is enough to focus on what happens at the edges of the interval. Redoing the calculations using the upper bounds of the two intervals,

²¹These calculations assume that the probability of a match if the suspect and the suspect's dog are the sources is one.

²²These are 99% credible intervals starting with uniform priors. A 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 Kruschke (2015).

.037 and .103, yields the following:

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

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 6). The calculation for the lower bounds, .015 and .002, yields the following:

$$P(\text{dog} \wedge \text{hair} | \neg \text{source}) = .015 \times .002 = .00003$$

This number is around 0.46 times lower than the original estimate.

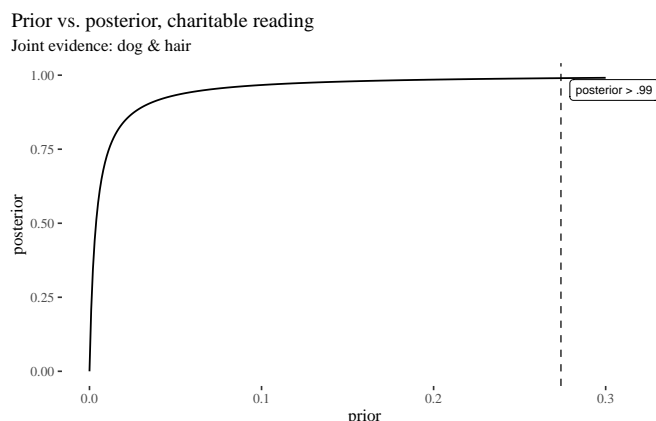


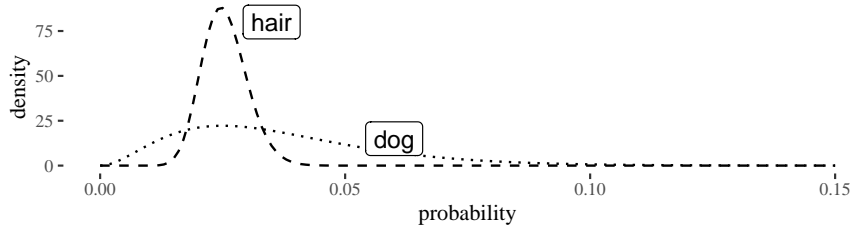
Figure 6: Impact of dog fur and human hair evidence on the prior, charitable reading.

Using plausible ranges for the match probabilities leaves the impression that any value in the interval is just as good as any other. But this impression will often misrepresent the data. Figure 7 (upper part) depicts the probability distribution of different match probabilities given the sample data—the actual number of matches found in the sample database. That is, 29 matches were found in a sample database of human hair of size 1148, and 2 matches were found in a sample database of dog fur of size 78. As expected, some random match probabilities are more likely than others. And since the sizes of the two databases are also different, the distributions have different spreads. The smaller the database the greater the spread, the greater the uncertainty about match probability. The trouble with imprecise probabilities is that these nuanced information is entirely lost.

Figure 7 (lower part) depicts the probability distribution for the joint match probability associated with both items of match evidence, hair and fur evidence.

Interestingly, the distribution for the joint evidence is not symmetric. This means that the most likely value of the joint match probability (and the bulk of the distribution, really) does not simply lie in the middle between the edges. Therefore, only relying on the edges can lead to overestimating or underestimating the probabilities at play.

Conditional densities for individual items of evidence if the source hypothesis is



Conditional density for joint evidence
(with .99 and .9 HPDIs)

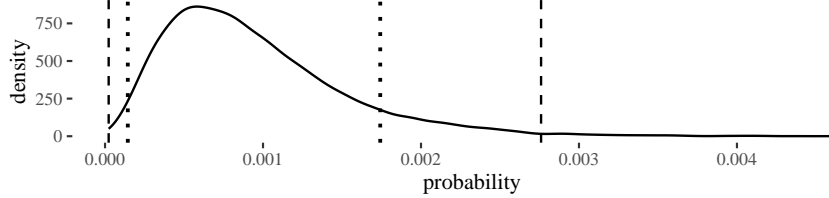


Figure 7: Beta densities for individual items of evidence and the resulting joint density with .99 and .9 highest posterior density intervals, assuming the sample sizes as discussed and independence, with uniform priors.

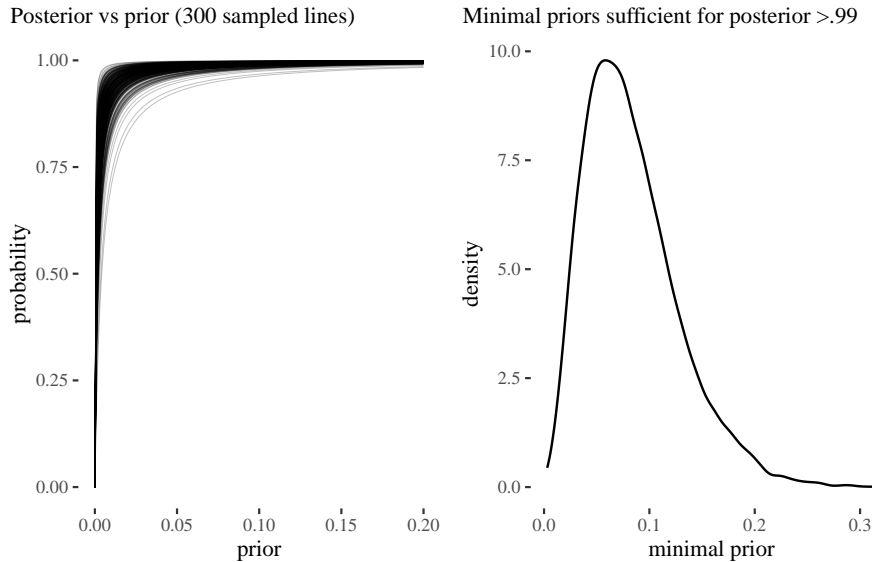


Figure 8: 300 lines illustrating the uncertainty about the dependence of the posterior on the prior given aleatory uncertainty about the evidence, with the distribution of the minimal priors required for the posterior to be above .99.

This, then, is the main claim illustrated in this section: higher-order approach to evidence evaluation is more reliable and more honest about the uncertainties involved. 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.

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 formal epistemology and in the legal scholarship and should be properly used to incorporate second-order uncertainties in evidence evaluation and incorporation.

7 Computational and representational considerations

The higher-order framework we are advocating is not only applicable to the evaluation of individual pieces of evidence. Complex bodies of evidence and hypotheses—for example, those often represented by Bayesian networks—can also be approached from this perspective. The general strategy is this: (1) capture the uncertainties involving the individual items of evidence in a modular fashion using the standard tools for statistical inference. (2) Elicit other probabilities or densities from experts²³, (3) put those together using a structure similar to that of a Bayesian network, except allowing for uncertainties of various levels to be put together—a usual tool for such a representation is a probabilistic program (Bingham et al., 2021), and (4) perform inference evaluating the relevant probabilities or densities of interest.

If the reader is more used to thinking in terms of Bayesian networks, a somewhat restrictive but fairly straightforward way to conceptualize a large class of such programs is to imagine a probabilistic program as stochastically generating Bayesian networks using our uncertainty about the parameter values, update with the evidence, and propagate uncertainty to approximate the marginal posterior for nodes of interest.

As an illustration, let us start with a simplified Bayesian network developed by Fenton & Neil (2018). The network is reproduced in Figure 9 and represents the key items of evidence in the infamous British case *R. v. Clark* (EWCA Crim 54, 2000).²⁴

In a Bayesian network the arrows depict direct relationships of influence between variables, and nodes—conditional on their parents—are taken to be independent of their non-descendants. A murder and B murder are binary nodes corresponding to whether Sally Clark's sons, call them A and B, were murdered. These nodes influence whether signs of disease (A disease and B disease) and bruising (A bruising and B bruising) were present. Also, since A's death preceded in time B's death, whether A was murdered casts some light on the probability that B was also murdered.

The choice of the probabilities in the network is quite specific, and it is not clear where such precise values come from. The standard response invokes *sensitivity analysis*: a range of plausible values is tested. As already discussed, this approach ignores the shape of the underlying distributions. Sensitivity analysis does not make any difference between probability measures (or point estimates) in terms of their plausibility, but some will be more plausible than others. Moreover, if the sensitivity analysis is guided by extreme values, these might play an undeservedly strong role. These concerns can be addressed, at least in part, by recourse to higher-order probabilities. In a precise Bayesian network, each node is associated with a probability table determined by a finite list of numbers (precise probabilities). But suppose that, instead of precise numbers, we have densities over parameter values for the numbers in the probability tables.²⁵ An example for the Sally Clark case is represented in Figure 10.

²³For expert elicitation of densities in a parametric fashion and the discussion of the improvement to which doing so instead of eliciting point values leads, see (O'Hagan et al., 2006).

²⁴Sally Clark's first son died in 1996 soon after birth, and her second son died in similar circumstances a few years later in 1998. At trial, the pediatrician Roy Meadow testified that the probability that a child from such a family would die of Sudden Infant Death Syndrome (SIDS) was 1 in 8,543. Meadow calculated that therefore the probability of both children dying of SIDS was approximately 1 in 73 million. Sally Clark was convicted of murdering her infant sons. The conviction was reversed on appeal. The case of appeal was based on new evidence: signs of a potentially lethal disease were found in one of the bodies.

²⁵The densities of interests can then be approximated by (1) sampling parameter values from the specified distributions, (2) plugging them into the construction of the BN, and (3) evaluating the probability of interest in that

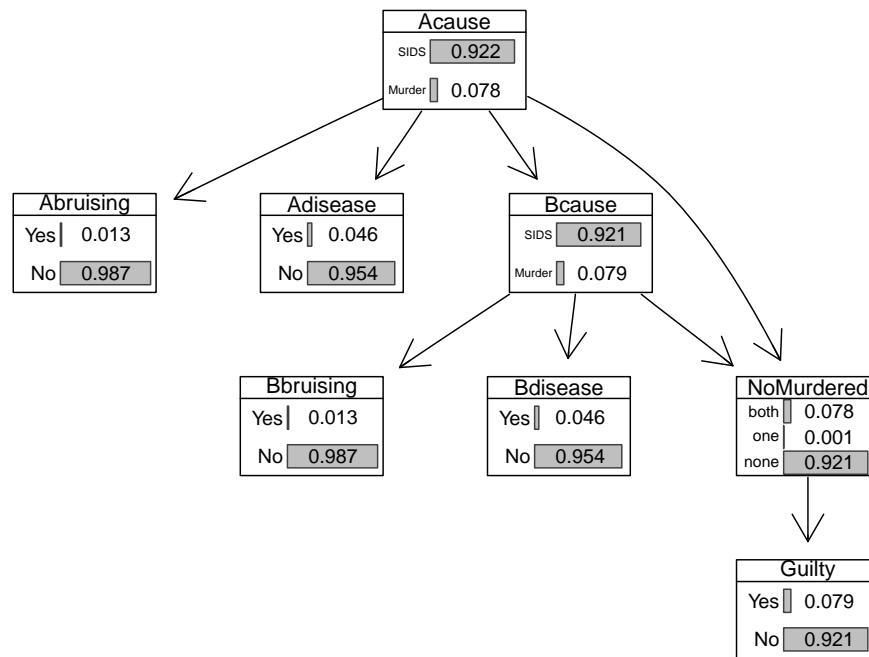


Figure 9: Bayesian network for the Sally Clark case, with marginal prior probabilities.

Using the probabilistic program, we can investigate the impact of different items of evidence on Sally Clark’s probability of guilt (Figure 10). The starting point is the prior density for the Guilt node (first graph). Next, the network is updated with evidence showing signs of bruising on both children (second graph). Next, the assumption that both children lack signs of potentially lethal disease is added (third graph). Finally, we consider the state of the evidence at the time of the appellate case: signs of bruising existed on both children, but signs of lethal disease were discovered only on the first child. Interestingly, in the strongest scenario against Sally Clark (third graph), the median of the posterior distribution is above .95, but the uncertainty around that median is still too wide to warrant a conviction.²⁶ This underscores the fact that relying on point estimates can lead to overconfidence. Paying attention to the higher-order uncertainty about the first-order probability can make a difference to trial decisions.

N: I am still searching for a good fix of that plot

nl: This plot is not referenced anywhere, should it be visible?

precise BN. The list of the probabilities thus obtained will approximate the density of interest. In what follows we will work with sample sizes of 10k.

²⁶The lower limit of the 89% Highest Posterior Density Intervals (HPDI) is at .83.

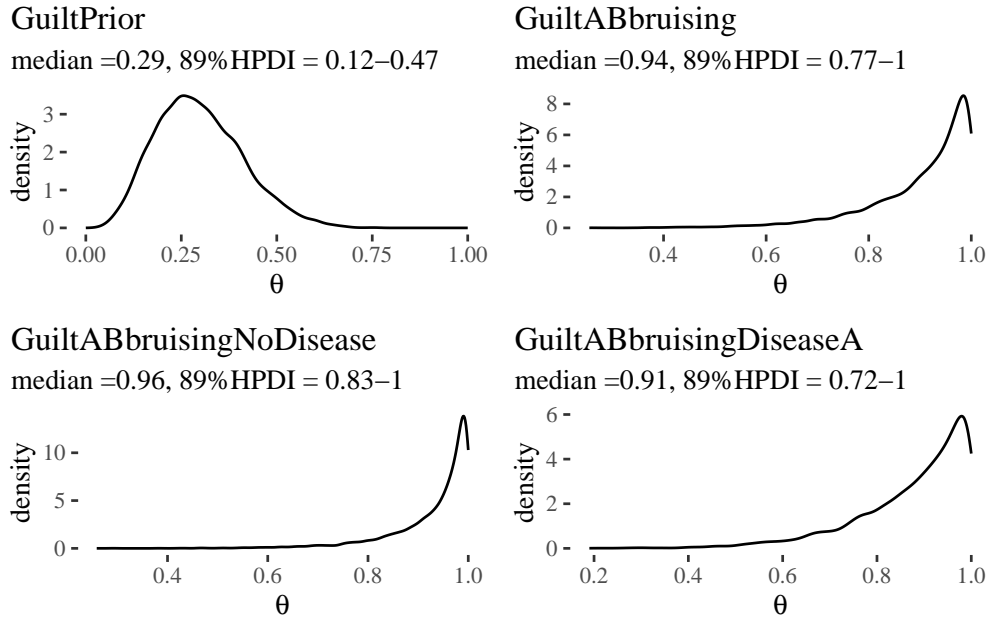


Figure 11: Impact of incoming evidence in the Sally Clark case.

One question that arises is how this approach relates to the standard method of using likelihood ratios to report the value of the evidence. On this approach, the conditional probabilities that are used in the likelihood ratio calculations are estimated and come in a package with an uncertainty about them. Accordingly, these uncertainties propagate: to estimate the likelihood ratio while keeping track of the uncertainty involved, we can sample probabilities from the selected distributions appropriate for the conditional probabilities needed for the calculations, then divide the corresponding samples, obtaining a sample of likelihood ratios, thus approximating the density capturing the recommended uncertainty about the likelihood ratio. Uncertainty about likelihood ratio is just propagated uncertainty about the involved conditional probabilities. For instance, we can use this tool to gauge our uncertainty about the likelihood ratios corresponding to the signs of bruising in son A and the presence of the symptoms of a potentially lethal disease in son A (Figure 12).

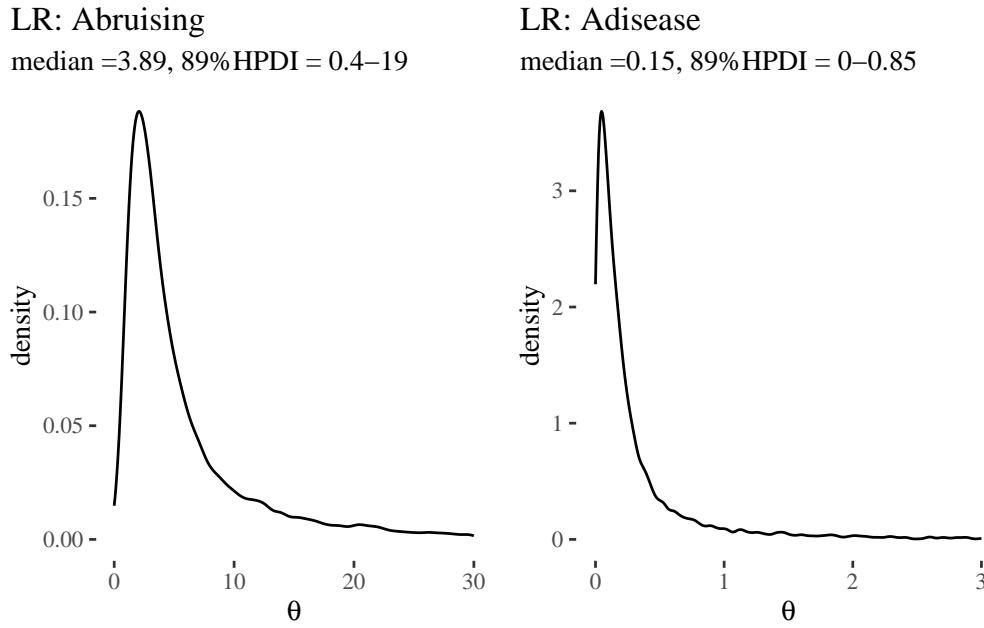


Figure 12: Likelihood ratios for bruising and signs of disease in child A in the Sally Clark case.

8 Discussion

Our approach does involve multiple parameters, uncertainty about them, along with a dependency structure between random variables. So it is only natural to ask whether what we propose is not just an old wolf in a new sheep’s clothing, as one might think that what looks like a DAG and quacks like a DAG is always a hierarchical model. In this section we briefly clarify what the answer to this question is.

First, we need some clarity on what a Bayesian hierarchical model is. In the widest sense of the word, these are mathematical descriptions involving multiple parameters such that credible values for some of them meaningfully depend on the values of other parameters, and that dependencies can be re-factored into a chain of dependencies. For instance, think about a joint parameter space for two parameters θ and ω , where $p(\theta, \omega | D) \propto p(D | \theta, \omega) p(\theta, \omega)$. If, further, some independence-motivated re-factoring of the right-hand side—for instance as $p(D | \theta) p(\theta | \omega) p(\omega)$ —is possible, we are dealing with a hierarchical model in the wide sense of the word.

Such models usually come useful when we are dealing with clustered data, such as a cohort study with repeated measures, or some natural groupings at different levels of analysis. Then, lower-level parameters are treated as i.i.d. and share the same parameter distribution characterized by some hyper-parameters in turn characterized by a prior distribution. As a simple example consider a scenario in which we are dealing with multiple coins created by one mint—each coin has its own bias θ_i , but also there is some commonality as to what these biases are in this mint, represented by a higher-level parameter θ . Continuing the example, assume $\theta_i \sim \text{Beta}(a, b)$ and $y_{i|s} \sim \text{Bern}(\theta_s)$, where the former distribution can be re-parametrized as $\text{Beta}(\omega(k-2) + 1, (1-\omega)(k-2) + 1)$. Let’s keep k fixed, ω is our expected value of the θ_i parameters, with some dispersion around it determined by k . Now, if we also are uncertain about ω and express our uncertainty about it in terms a density $p(\omega)$, we got ourselves a hierarchical model with joint prior distribution over parameters $\prod p(\theta_i | \omega) p(\omega)$.

As another example, one can develop a multilevel regression model of the distributions of the random levels in various counties, where both the intercept and the slope vary with counties

by taking

$y_i \sim \text{Norm}(\alpha_{j[i]} + \beta_{j[i]}x_i, \sigma_y^2)$, where j is a county index, $\alpha_j \sim \text{Norm}(\mu_\alpha, \sigma_\alpha^2)$, and $\beta_j \sim \text{Norm}(\mu_\beta, \sigma_\beta^2)$. Then, running the regression one estimates both the county-level coefficients, and the higher-level parameters.

Our approach is similar to the standard hierarchical models in the most general sense: there is a meaningful independence structure and distributions over parameter values that we are working with. However, our approach is unlike such models in a few respects. For one, we are not dealing with clustered data, and the random variables are mostly propositions and their truth values. Given a hypothesis H and an item of evidence E for it, there seems to be no interesting conceptualization on which the underlying data would be clustered. For example, considering stains at a crime scene as a subgroup of crimes being committed doesn't make logical sense. Yes, there is dependency between these phenomena, but describing it as clustering would be at least misleading. Second, the dependencies proceed through the values of the random variables which are **not** parameters, but rather truth-values, and require also conditional uncertainties regarding the dependencies between these truth-values.

Again, continuing the hypothesis-evidence example, we have $H \sim \text{Bern}(p_h)$, $p_h \sim \text{Beta}(a_h, b_h)$, and $E \sim \text{Bern}(p_e)$. But then we also have the beta distributions for the probability of the evidence conditional on the actual values of the random variables—the truth-values—thus $p_e|H = 1 \sim \text{Beta}(a_+, b_+)$ and $p_e|H = 0 \sim \text{Beta}(a_-, b_-)$. But the re-factoring in terms of the actual values of the random variables (which just happen to resemble probabilities because they are truth values) makes it quite specific, at the same time allowing for the computational use of a probabilistic program. Finally, the reasoning we describe is not a regression the way it is normally performed: the learning task is delegated to the bottom level of whatever happens to the Bayesian networks once updated with evidence. We would prefer to reserve the term *hierarchical model* for a class of models dealing with interesting cluster structures in the data. A more fitting term for the representation tool we propose should be used here is *probabilistic programs*. We do not claim any originality in devising this tool: it's an already existing tool. What we argue for, though, is its ability for being usefully deployed in the context of forensic evidence evaluation and integration with other assumptions and hypotheses.

Perhaps, you might dislike the idea of going higher-order for theoretical reasons. One might be that you don't like the complexity. This seems to be the line taken by Bradley, who refuses to go higher-order for the following reason:

Why is sets of probabilities the right level to stop the regress at? Why not sets of sets? Why not second-order probabilities? Why not single probability functions? This is something of a pragmatic choice. The further we allow this regress to continue, the harder it is to deal with these belief representing objects. So let's not go further than we need. 131-132

We have argued extensively, that given the difficulties of both PP and IP and how the current approach handles it, we are not going further than we need in using higher-order probabilities. We're going where we should be. And the supposed pragmatic concerns that one might have are unclear: parameter uncertainty, approximations and other computational methods I have used in fact quite embedded in Bayesian statistical practice and decent computational tools for the framework I propose are available.²⁷

Another concern that you might have is that it is not clear what the semantics of such an

²⁷Also, you can insist that instead of going higher order we could just take our sample space to be the cartesian product of the original sample space and parameter space, or use parameters having certain values as potential states of a bayesian network. If you prefer not to call such approaches first-order, I don't mind, as long as you effectively end up assigning probabilities to certain probabilities, the representation means I discussed in this paper should be in principle available to you.

approach should look like. While a more elaborate account is beyond the scope of this paper, the general gist of the approach can be modeled by a slight modification of a framework of probabilistic frames (Dorst, 2022b, 2022a). Start with a set of possible worlds W . Suppose you consider a class of probability distributions D , a finite list of atomic sentences q_1, \dots, q_2 corresponding to subsets of W , and a selection of true probability hypotheses C (think of the latter as omniscient distributions, $C \subseteq D$, but in principle this restriction can be dropped if need be). Each possible world $w \in W$ and a proposition $p \subseteq W$ come with their true probability distribution, $C_{w,p} \in D$ corresponding to the true probability of p in w , and the distribution that the expert assigns to p in w , $P_{w,p} \in D$. Then, various propositions involving distributions can be seen as sets of possible worlds, for instance, the proposition that the expert assigns d to p is the set of worlds w such that $P_{w,p} = d$.²⁸

Added this passage

Cite <https://arxiv.org/abs/2302.09656>

The reader might also be eager to point out that some of the techniques we mentioned are already available to the impreciser. For instance, one can use uniform sampling with Bayesian networks to approximate the impreciser's epistemic commitments given their assumptions. This doesn't mathematically differ from relying on probabilistic programs with the restriction that some variables corresponding to probabilities are sampled from distributions that are uniform over a set of values determined by the impreciser's representor. However, this only shows that a computational implementation of the impreciser's perspective is not out of reach—which is to be expected. A critical survey of an approach along these lines uses roughly this computational approach to argue that in complex reasoning situations, if the impreciser's stance is taken, "the imprecision of inferences increases rapidly as new premises are added to an argument". This is in line with our criticism.

add ref to <https://www.sciencedirect.com/science/article/pii/S000437029600021>

Appendix: propriety

Appendix: the strict propriety of $\mathcal{J}_{D_{KL}}^2$

Let us start with a definition.

Definition 1 (concavity). *A function f is convex over an interval (a, b) just in case for all $x_1, x_2 \in (a, b)$ and $0 \leq \lambda \leq 1$ we have:*

$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2)$$

A function f is concave just in case:

$$f(\lambda x_1 + (1 - \lambda)x_2) \geq \lambda f(x_1) + (1 - \lambda)f(x_2)$$

A function f is strictly concave just in case the equality holds only if either $\lambda = 0$ or $\lambda = 1$.

For us it is important that if a function is twice differentiable on an interval, then it is (strictly) concave just in case its second derivative is non-positive (negative). In particular, as $(\log_2(x))'' = -\frac{1}{x^2 \ln(2)}$, \log_2 is strictly concave over its domain.²⁹

Lemma 1 (Jensen's inequality). *If f is concave, and g is any function of a random variable, $\mathbb{E}(f(g(x))) \leq f(\mathbb{E}(g(x)))$. If f is strictly concave, the equality holds only if $g(x) = \mathbb{E}g(x)$, that is, if $g(x)$ is constant everywhere.*

²⁸There is at least one important difference between this approach and that developed by Dorst. His framework is untyped, which allows for an enlightening discussion of the principle of reflection and alternatives to it. In this paper I prefer to keep this complexity apart and use an explicitly typed set-up.

²⁹I line with the rest of the paper, we'll work with log base 2. We could equally well use any other basis.

Proof. For the base case consider a two-point mass probability function. Then,

$$p_1 f(g(x_1)) + p_2 f(g(x_2)) \leq f(p_1 g(x_1) + p_2 g(x_2))$$

follows directly from the definition of concavity, if we take $\lambda = p_1$, $(1 - \lambda) = p_2$, and substitute $g(x_1)$ and $g(x_2)$ for x_1 and x_2 .

Now, suppose that $p_1 f(g(x_1)) + p_2 f(g(x_2)) = f(p_1 g(x_1) + p_2 g(x_2))$ and that f is strictly concave. That means either $(p_1 = 1 \wedge p_2 = 0)$, or $(p_1 = 0 \wedge p_2 = 1)$. Then either x always takes value x_1 , in the former case, or always takes value x_2 , in the latter case. $\mathbb{E}g(x) = p_1 g(x_1) + p_2 g(x_2)$, which equals $g(x_1)$ in the former case and $g(x_2)$ in the latter.

Now suppose Jensen's inequality and the consequence of strict concavity holds for $k - 1$ mass points. Write $p'_i = \frac{p_i}{1 - p_k}$ for $i = 1, 2, \dots, k - 1$. We now reason:

$$\begin{aligned} \sum_{i=1}^k p_i f(g(x_i)) &= p_k f(g(x_k)) + (1 - p_k) \sum_{i=1}^{k-1} p'_i f(g(x_i)) \\ &\leq p_k f(g(x_k)) + (1 - p_k) f\left(\sum_{i=1}^{k-1} p'_i g(x_i)\right) \quad \text{by the induction hypothesis} \\ &\leq f\left(p_k g(x_k) + (1 - p_k) \sum_{i=1}^{k-1} p'_i g(x_i)\right) \quad \text{by the base case} \\ &= f\left(\sum_{i=1}^k p_i g(x_i)\right) \end{aligned}$$

Notice also that at the induction hypothesis application stage we know that the equality holds only if $p_k = 1 \vee p + k = 0$. In the former case $g(x)$ always takes value $x_k = \mathbb{E}g(x)$. In the latter case, p_k can be safely ignored and $\sum_{i=1}^k p_i g(x_i) = \sum_{i=1}^{k-1} p'_i g(x_i)$ and by the induction hypothesis we already know that $\mathbb{E}g(x) = g(x)$. □

In particular, the claim holds if we take $g(x)$ to be $\frac{q(x)}{p(x)}$ (were both p and q are probability mass functions), and f to be \log_2 . Then, given that A is the support set of p , we have:

$$\sum_{x \in A} p(x) \log_2 \frac{q(x)}{p(x)} \leq \log_2 \sum_{x \in A} p(x) \frac{q(x)}{p(x)}$$

Moreover, the equality holds only if $\frac{q(x)}{p(x)}$ is constant, that is, only if p and q are the same pmfs. Let's use this in the proof of the following lemma.

Lemma 2 (Information inequality). *For two probability mass functions p, q , $D_{\text{KL}}(p, q) \geq 0$ with equality iff $p = q$.*

Proof. Let A be the support set of p , and let q be a probability mass function whose support

is B .

$$\begin{aligned}
-D_{\text{KL}}(p, q) &= -\sum_{x \in A} p(x) \log_2 \frac{p(x)}{q(x)} && \text{(by definition)} \\
&= \sum_{x \in A} p(x) - (\log_2 p(x) - \log_2 q(x)) \\
&= \sum_{x \in A} p(x) (\log_2 q(x) - \log_2 p(x)) \\
&= \sum_{x \in A} p(x) \log_2 \frac{q(x)}{p(x)} \\
&\leq \log_2 \sum_{x \in A} p(x) \frac{q(x)}{p(x)} && \text{by Jensen's inequality} \\
&\text{(and the equality holds only if } p = q) \\
&= \log_2 \sum_{x \in A} q(x) \\
&\leq \log_2 \sum_{x \in B} q(x) \\
&= \log(1) = 0
\end{aligned}$$

□

Observe now that D_{KL} can be decomposed in terms of cross-entropy and entropy.

Lemma 3 (decomposition). $D_{\text{KL}} = H(p, q) - H(p)$.

Proof.

$$\begin{aligned}
D_{\text{KL}}(p, q) &= \sum_{p_i} (\log_2 p_i - \log_2 q_i) \\
&= -\sum_{p_i} (\log_2 q_i - \log_2 p_i) \\
&= -\sum_{p_i} \log_2 q_i - \sum_{p_i} -\log_2 p_i \\
&= \underbrace{-\sum_{p_i} \log_2 q_i}_{H(p, q)} - \underbrace{\sum_{p_i} -\log_2 p_i}_{H(p)}
\end{aligned}$$

□

With information inequality this easily entails Gibbs' inequality:

Lemma 4 (Gibbs' inequality). $H(p, q) \geq H(p)$ with identity only if $p = q$.

Now we are done with our theoretical set-up. Here is how it entails the propriety of $\mathcal{J}_{D_{\text{KL}}}^2$. First, let's systematize the notation. Consider a discretization of the parameter space $[0, 1]$ into n equally spaced values $\theta_1, \dots, \theta_n$. For each i the “true” second-order distribution if the true parameter indeed is θ_i —we'll call it the indicator of θ_i —is defined by

$$\text{Ind}^k(\theta_i) = \begin{cases} 1 & \text{if } \theta_i = \theta_k \\ 0 & \text{otherwise} \end{cases}$$

I will write Ind_i^k instead of $Ind^k(\theta_i)$.

Now consider a probability distribution p over this parameter space, assigning probabilities p_1, \dots, p_n to $\theta_1, \dots, \theta_n$ respectively. It is to be evaluated in terms of inaccuracy from the perspective of a given ‘true’ value θ_k . The inaccuracy of p if θ_k is the ‘true’ value, is the divergence between Ind^k and p .

$$\begin{aligned} \mathcal{J}_{D_{KL}}^2(p, \theta_k) &= D_{KL}(Ind^k, p) \\ &= \sum_{i=1}^n Ind_i^k (\log_2 Ind_i^k - \log_2 p_i) \end{aligned}$$

Note now that for $j \neq k$ we have $Ind_j^k = 0$ and so $Ind_j^k (\log_2 Ind_j^k - \log_2 p_j) = 0$. Therefore we continue:

$$= Ind_k^k (\log_2 Ind_k^k - \log_2 p_k)$$

Further, $Ind_k^k = 1$ and therefore $\log_2 Ind_k^k = 0$, so we simplify:

$$= -\log_2 p_k$$

Now, let’s think about expected values. First, what the inaccuracy of p as expected by p , $EI(p, p)$?

$$\begin{aligned} EI(p, p) &= \sum_{i=1}^n p_i \mathcal{J}_{D_{KL}}^2(p, \theta_k) \\ &= \sum_{i=1}^n p_i (-\log_2 p_k) \\ &= -\sum_{i=1}^n p_i \log_2 p_k = H(p) \end{aligned}$$

Analogously, the inaccuracy of q as expected from the perspective of p is:

$$\begin{aligned} EI(p, q) &= \sum_{i=1}^n p_i (-\log_2 q_i) \\ &= -\sum_{i=1}^n p_i \log_2 q_i = H(p, q) \end{aligned}$$

But that means, by Gibbs’ inequality, that $EI(p, q) \geq EI(p, p)$ unless $p = q$, which completes the proof.

References

- Bingham, E., Koppel, J., Lew, A., Ness, R., Tavares, Z., Witty, S., & Zucker, J. (2021). Causal probabilistic programming without tears. *Proceedings of the Third Conference on Probabilistic Programming*.
- 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.
- Deadman, H. A. (1984b). Fiber evidence and the wayne williams trial (part i). *FBI L. Enforcement Bull.*, 53, 12–20.
- Dorst, K. (2022a). Higher-order evidence. In M. Lasonen-Aarnio & C. Littlejohn (Eds.), *The routledge handbook for the philosophy of evidence*. Routledge.
- Dorst, K. (2022b). Higher-order uncertainty. In M. Skipper & A. S. Petersen (Eds.), *Higher-order evidence: New essays*.
- Elkin, L. (2017). *Imprecise probability in epistemology* (PhD thesis). Ludwig-Maximilians-Universität; Ludwig-Maximilians-Universität München.
- Fenton, N., & Neil, M. (2018). *Risk assessment and decision analysis with bayesian networks*. Chapman; Hall.
- 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>
- Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477), 359–378. <https://doi.org/10.1198/016214506000001437>
- Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15(5), 559–570. [https://doi.org/10.1175/1520-0434\(2000\)015%3C0559:DOTCRP%3E2.0.CO;2](https://doi.org/10.1175/1520-0434(2000)015%3C0559:DOTCRP%3E2.0.CO;2)
- 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.
- Kruschke, J. (2015). *Doing bayesian data analysis (second edition)*. Boston: Academic Press.
- 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>
- O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., ... Rakow, T. (2006). *Uncertain judgements: Eliciting experts' probabilities*.
- Pettigrew, R. (2012). *Epistemic utility and norms for credences*.
- 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>
- 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.

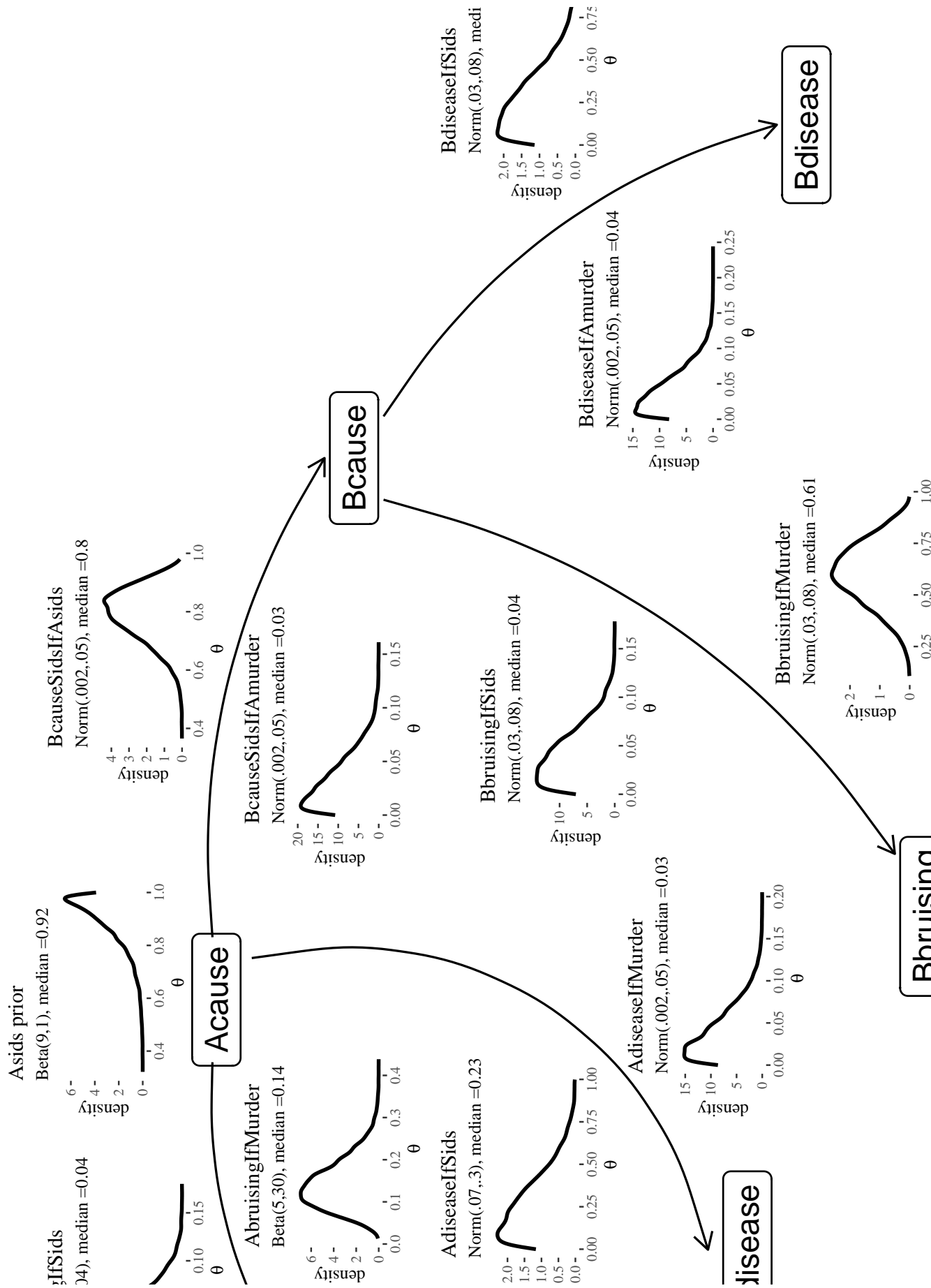


Figure 10: An illustration of a probabilistic program for the Sally Clark case.