Seamus Bradley*

Learning by Ignoring the Most Wrong

<https://doi.org/10.1515/krt-2021-0028>

Published online December 24, 2021

Abstract: Imprecise probabilities (IP) are an increasingly popular way of reasoning about rational credence. However they are subject to an apparent failure to display convincing inductive learning. This paper demonstrates that a small modification to the update rule for IP allows us to overcome this problem, albeit at the cost of satisfying only a weaker concept of coherence.

Keywords: imprecise probability, formal epistemology, probability theory

1 Imprecise Probabilities

Consider representing your rational state of belief – your credal state – by a *set* of probability functions, $\mathbb{P}$, with a common algebra of events $\mathcal{A}$. Call $\mathbb{P}$ a *credal set*.¹ We might interpret this as a set of possible sharpenings of a somewhat indeterminate credal state, for example.² Define $\mathbb{P}(X) = \{\mathbf{pr}(X), \mathbf{pr} \in \mathbb{P}\}$, the set-valued function defined over the same algebra as the members of $\mathbb{P}$; and define $\overline{\mathbb{P}}(X) = \sup \mathbb{P}(X)$, the “upper probability” that takes the largest value assigned to an event X by members of $\mathbb{P}$. We can also define a “lower probability”: $\underline{\mathbb{P}}(X) = \inf \mathbb{P}(X)$. This is a simple imprecise probability (IP) theory of rational belief.

We’re going to follow the standard practice of exploring the rationality of a credal state by considering what consequences it has for betting behaviour. Let’s

¹ Despite its naturalness and familiarity, it should be noted that the framework of credal sets and lower probabilities is not as general or expressive as some other IP models, such as lower previsions or sets of desirable gambles. For our purposes, this extra representational power will not be necessary.

² For more on interpreting IP and IP theory more generally, see Augustin et al. (2014); Bradley (2019); Joyce (2005, 2010), Levi (1974, 1980); Walley (1991).

Many thanks to the audience in Munich (online) for their comments. This project has received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No 792292.

***Corresponding author: Seamus Bradley**, University of Leeds, Woodhouse Lane, Leeds LS2 9JT, London, E-mail: scmb Bradley@gmail.com

make the assumption that if your credal state is $\mathbb{P}$, then $\mathbb{P}$ serves as your “maximum buying price” for gambles, meaning that $\mathbb{P}$ fills the role of q below. Call the bet “win $1 - x$ if A , lose x otherwise” a “unit bet on A (with a price of x)”.

One-Sided Betting

A unit bet on A with price x is acceptable if $q(A) > x$.

If you accept a unit bet on A , then the bookie has accepted a unit bet on $\neg A$ with price $1 - x$. Call the function q that plays the role of determining the acceptability of bets your “betting quotient”. The standard practice we are following here is to identify your betting quotient with your credence, but they are conceptually distinct.

If your credences inform your betting behaviour in this way, then it seems a minimal requirement of rationality is that your betting behaviour should be such that you are not subject to a sure loss. It is a bad thing to be such that, whatever happens, you lose money.

Avoid Sure Loss

Your prices for gambles in a one-sided betting scenario should be such that no combination of acceptable bets is guaranteed to yield a loss in every state.

There is a further property that we might also find desirable, namely that any collection of bets that always wins you money is acceptable to you. This is what Hájek (2008) calls a “Czech book”. So we might want to add in this extra clause.

Coherence

Your prices for gambles in a one-sided betting scenario should be such that no combination of acceptable bets is guaranteed to yield a loss in every state and every combination of bets that yield a gain in every state is acceptable.

Coherence is strictly stronger than avoiding sure loss. We then have the following result: you are coherent if and only if your buying betting quotients are the lower probabilities for some credal set (Troffaes and de Cooman 2014, Theorem 4.38).³

³ Note this result is for the separate coherence of an individual lower probability, not the joint coherence of a prior and conditional lower probability. Results there are somewhat more complex.

Note that the standard precise probability theory is recovered if we require more from your betting quotient:

Two-Sided Betting

- A unit bet on A with price x is acceptable if $q(A) > x$.
- If the above bet is not acceptable then a unit bet on $\neg A$ with price $1 - y$ is acceptable for any $y > x$.

Satisfying coherence when your betting quotient function also determines what bets you sell requires that $\underline{\mathbb{P}}(X) = \overline{\mathbb{P}}(X)$ for all X ; in other words you have to have precise probabilistic prices. This is what is standardly called the “Dutch book argument”. Getting from here to an argument that your rational credences ought to have a certain form requires some kind of premise that connects your rational prices for bets to your credences. Since my main focus is elsewhere, I’m just going to stipulate that we have such a principle, and I’ll talk about prices and credences interchangeably.

There’s a few moving parts, here, and I think it will be helpful to be clear about the scope of the argument I’m making, so let’s run through it. Rational agents have subjective credences; we’re being deliberately non-committal about what these are or what structure they have. Credences determine betting quotients; this is a standard move, and one I’m not questioning or really discussing much.⁴ If credences are probabilities, then the standard assumption is to identify credences and betting quotients. An alternative would be to argue that credences *together with a risk profile* determine betting quotients (Buchak 2013), but we won’t pursue that here. Betting quotients determine acceptability of bets; this is what the **One-sided betting** condition does. For our purposes, we’re only really interested in unit bets on propositions, but that is merely for presentational convenience. The full theory of lower previsions (Augustin et al. 2014; Troffaes and de Cooman 2014) determines the acceptability for arbitrary gambles on a space of possibilities. Rationality conditions often take the form of saying “finding bets of this form acceptable is irrational” or “failing to find bets of this form acceptable is irrational”: **Avoid sure loss** is clearly of the former form and the additional clause in **Coherence** is of the latter. Finally acceptability of a bet is sufficient but perhaps not necessary for being choiceworthy. “Choiceworthy” here is an adjective that applies to all and only the actions sanctioned by your decision theory. A decision theory is typically thought of as whatever it is that gets you from your credences

⁴ I am here using the slightly non-standard terminology of “betting quotient” because in the mathematical literature, these objects get called “lower previsions” or, in our circumscribed case “lower probabilities”, but I want to emphasize the conceptual difference between a credence and the betting quotient it determines.

and your utilities (representing your beliefs and desires) to judgements about what you ought to choose. So, to recap, I am being partly agnostic about what decision theory you use, so long as that decision theory is at least compatible with endorsing the acceptability judgements of a straightforward (risk neutral) application of betting quotients. For example, most of the decision rules that Bradley (2015) discusses will endorse the acceptability judgements of **One-sided betting** (since the differences between the rules mostly show up in choices involving more than two options).⁵ The final complication is the constraints on combinations of bets. In this paper I take the view that if you find gamble g acceptable and you find gamble g' acceptable, then you find the combination of g and g' acceptable. This is standard practice and a necessary part of the Dutch book theorem. In a context of sequential decision making Rinard (2015) and Bradley and Steele (2014a) argue against this “package principle”. We shall return to this point later.

2 Belief Inertia

We can use the same betting methodology that we used to provide structure to rational credences, to constrain the rational response to evidence. Let's start with the precise probabilities case. Coherence should not just constrain attitudes at a time, but also attitudes over time. Consider your attitudes before updating on some proposition E , and after having done the update. Call these $\mathbf{pr}$ and $\mathbf{pr}'$ respectively. It is not enough for $\mathbf{pr}$ and $\mathbf{pr}'$ to both satisfy the axioms of probability individually: they should also be such that you can't incur a sure loss in moving from one to the other on learning E . In short, the two credences should be *jointly coherent* and not merely separately coherent. In practice, what this means is that $\mathbf{pr}'(X) = \mathbf{pr}(X|E) = \frac{\mathbf{pr}(XE)}{\mathbf{pr}(E)}$.⁶ Now the question is: does updating in this way mean that we acquire better credences as we accrue evidence?⁷ As we'll see below, the answer is: “sometimes”.

In this paper, we're going to focus on a particular class of credence functions that do seem to exhibit reasonable inductive learning behaviour. Consider a precise credence function that is determined by two parameters, $\gamma \in (0, 1)$ and $\lambda > 0$ where the credence that the next coin will land heads is $\mathbf{pr}(H) = \gamma$, and learning

⁵ The decision problems here are all binary choices between accepting the bet and the status quo.

⁶ For the purposes of this essay, I am interested in learning E with certainty only, so I will not be discussing Jeffrey conditioning or any other form of update.

⁷ “Better” in what sense? We'll discuss this briefly at the end of this section.

that there were h heads and t tails in the last $n = h + t$ tosses of the coin (call this evidence $S_n = h$) gives you an updated credence of $\mathbf{pr}(H|S_n = h) = \frac{\gamma\lambda + S_n}{\lambda + n}$.⁸ Any such credence is jointly coherent for all learning propositions $S_n = h$. We shall call such probability functions “learning priors”.⁹ Note that Carnap’s continuum of inductive methods (for one predicate) are the special case when $\gamma = \frac{1}{2}$.

Each $\mathbf{pr}$ is identified by two parameters γ and λ that capture a prior attitude to the bias of the coin or the chance of heads, and also how that prior responds to evidence. We can see this by rearranging the above expression to obtain:

$$\mathbf{pr}(H|S_n = h) = \left(\frac{n}{n + \lambda}\right)\frac{S_n}{n} + \left(\frac{\lambda}{n + \lambda}\right)\gamma \quad (1)$$

γ determines the prior attitude to the next toss’ landing heads, and λ governs how much your attitude changes in response to evidence about previous tosses. So if we keep λ fixed and vary γ , we have a range of priors that start at different points, but learn at roughly the same rate (Figure 1).¹⁰ On the other hand, if we fix γ and vary λ , we have a range of priors that start with the same prior credence for H , but move different amounts on the basis of the same evidence (Figure 2). Note also that as these priors acquire evidence, they become less prone to change on the basis of further evidence.

Given enough evidence, any such learning prior $\mathbf{pr}$ will converge on the true value of the chance with increasing evidence.¹¹ By the (strong) law of large numbers, as we increase the number of coin tosses (i.e. as n increases), the ratio $\frac{S_n}{n}$ will tend to the true chance, and as n gets larger, $\frac{\gamma\lambda + S_n}{\lambda + n}$ gets closer to $\frac{S_n}{n}$, since the right-hand term of (1) will get smaller as n increases. So each learning prior will converge on the true chance of heads as evidence increases. This seems like a good result. At least for this class of priors, we have sensible inductive inference behaviour.

8 Here and throughout H is a shorthand for $H_{n+1} = 1$: that is, the next toss of the coin will land heads, and $S_n = \sum^n H_i$.

9 These learning priors are expectations of beta distributions, but this detail plays no role in the following discussion.

10 A note on how to read the graphs. Each line represents a different learning prior. The x axis is labelled with the new evidence you receive at each time step. So you are first told that the first four tosses were all heads, then that the next four tosses were one head and three tails and so forth. The y axis is the probability of heads for the next toss. Source code for all the graphs in this paper is available here: <https://github.com/scmbradley/betalearn>.

11 Here and throughout, when I make some claim that x will converge to y , I mean that x will converge to y *almost surely*.

Figure 1: Learning priors for fixed λ .

Figure 2: Learning priors for fixed γ .

To be clear, the property that priors of this kind have is the following:

Convergence to the True Chance

For a sequence of i.i.d. binary random variables $H_i \in \{0, 1\}$, generated by probability $\mathbf{pr}^*(H) = p$, let $S_n = \sum_{i=1}^n H_i$. For almost all $(\mathbf{pr}^*)$ sequences of outcomes, $\mathbf{pr}(H|S_n) \rightarrow p$ as $n \rightarrow \infty$.

That is, you are almost certain to receive a sequence of outcomes such that you converge to having correct beliefs about the next coin toss.¹² The chance that you receive evidence that doesn't prompt you to converge on the truth is zero. More generally, if the desideratum is to approach some expert probability $(\mathbf{pr}^*)$, we want to converge (almost surely) to $\mathbf{pr}^*(H|S_n)$. Since coin tosses are independent of past tosses, $\mathbf{pr}^*(H) = \mathbf{pr}^*(H|S_n)$ for the chance function.

Let's turn to IP now. Joint coherence of an unconditional and a conditional lower probability is a little trickier than the precise case¹³ but for our purposes it is enough to say that generalised conditioning is the most informative coherent way to update: define a conditional credal set: $\mathbb{P}(-|E) = \{\mathbf{pr}(-|E), \mathbf{pr} \in \mathbb{P}, \mathbf{pr}(E) > 0\}$, the set obtained by conditioning each member of $\mathbb{P}$ on the same evidence. Note the caveat that we throw out any $\mathbf{pr}$ that assign zero probability to the evidence received. We shall return to that feature shortly.

Let's return to the coin-tossing problem to see how credal sets get stuck. Let $\mathbb{B}$ be the set of all learning priors. That is, $\mathbb{B}$ is a set containing priors, all of which can learn on the basis of evidence.

Suppose we've gathered evidence $S_n = h$. The conditional probabilities of H given S_n are all the numbers of the form: $\frac{\gamma\lambda+h}{\lambda+n}$. Hold h and n fixed, and for any $\lambda > 0$ and $\gamma \in (0, 1)$, that fraction is a possible probability value for heads. As λ gets bigger, this fraction tends to γ , so since γ is unconstrained, the range of values for the probability of heads, given $S_n = h$, covers all of $(0, 1)$ regardless of what evidence you acquire. The problem is that, even though each prior is learning, there are always more and more recalcitrant priors in $\mathbb{B}$ that cover the whole range of

¹² One might wonder why I am here focusing on “almost all” with respect to $\mathbf{pr}^*$ – the chance function – rather than with respect to your credence $\mathbf{pr}$. There are two reasons. One, when it comes to IP, it's easier to talk about “almost all” with respect to the chance than to talk about “almost all” with respect to each $\mathbf{pr} \in \mathbb{P}$. Two, I want my credences to fail to converge only for sequences of evidence that *actually* have vanishingly small chance, rather than for sequences of evidence that *I believe* to have vanishingly small chance.

¹³ See Augustin et al. (2014) Section 2.3, Troffaes and de Cooman (2014) chapter 13 and Zaffalon and Miranda (2013). What we call “Generalised Conditioning” is sometimes referred to as “Regular Extension”.

possible chances. The bigger λ is the less effect the evidence has on moving that prior towards the true chance.

There are a number of ways we could diagnose the problem that this example highlights. One problem is that $\mathbb{P}(H|S_n)$ fails to converge to $\mathbf{pr}^*(H)$. Indeed in the example just given, $\mathbb{P}(H|S_n) = (0, 1)$ for every possible (finite) sequence of evidence about coin flips. Alternatively, we could point to the fact that $\overline{\mathbb{P}}(H|S_n) - \underline{\mathbb{P}}(H|S_n)$ does not shrink as n increases. Call the former *failure to converge* and the latter *failure to narrow*. Convergence and narrowing are both desirable features of IP updating.

There are a number of ways IP updating can go that appear pathological. A credal set might be such that $\mathbb{P}(H|S_n)$ converges to something other than $\mathbf{pr}^*(H|S_n)$. Or $\mathbb{P}(H|S_n)$ might just fail to converge at all. Or $\mathbb{P}$ might converge but $\overline{\mathbb{P}}$ does not or vice versa. A credal set might fail to narrow by remaining stuck at $\mathbb{P}(H|S_n) = (0, 1)$ for all n (as in the above example) or $\mathbb{P}(H|S_n)$ might shrink somewhat to some non-trivial interval. Call the former *complete failure to narrow* and the latter *partial failure to narrow*. In this latter case, we might further want to distinguish cases where the interval contains $\mathbf{pr}^*$ and cases where it does not, with the former case seeming less bad somehow. These are not all logically distinct, obviously: complete failure to narrow precludes any sort of convergence, for example.

“Belief inertia” is used to refer to failure to narrow, but it’s often unclear or unspecified whether partial failure to narrow counts as inertia. It is also worth emphasising that solving failure to narrow is pretty easy,¹⁴ the tricky thing is doing so in a way that permits convergence. The problem for IP is not belief inertia per se, but rather a failure to learn inductively. Solving belief inertia won’t necessarily solve this problem: there are bad ways to make credal sets narrow. “Belief inertia” appears to sometimes be used as a synonym for “failure to learn inductively”, but I hope the above comments are enough to make clear that this is not correct.

So, even in quite friendly circumstances – every prior in the set can learn – certain IP prior credal sets are unable to learn from evidence. Levi (1980) noticed this problem (chapter 13), and Walley (1991) was also well aware of the issue. Philosophical interest in the topic has picked up recently, with a number of papers discussing it (Castro and Hart 2019; Joyce 2010; Lassiter 2020; Lyon 2017; Moss n.d.; Rinard 2013; Vallinder 2018). Belief inertia is normally brought up by these authors as a reason to be sceptical that IP can provide a reasonable theory of rational belief, or at least as a serious problem that a theory of IP ought to overcome. Since sensible inductive inference seems like an absolutely central requirement for such a theory,

¹⁴ For example, update by removing all priors that are further away from the midpoint than $1/n$ for some definition of midpoint, and some method of measuring distance.

we are owed an explanation of how to deal with belief inertia. I propose we address this problem by modifying the update rule for IP slightly.

3 Learning by Ignoring the Most Wrong

Recall that conditionalisation for a precise credence is defined, as is standard, by the following equation:

$$\mathbf{pr}(X|E) = \frac{\mathbf{pr}(XE)}{\mathbf{pr}(E)}.$$

Standard precise conditioning typically comes with the caveat that it only works for $\mathbf{pr}(E) > 0$. That is, if the evidence you receive has zero prior probability, conditioning doesn't work.¹⁵ There is a mathematical reason for this: you can't divide by zero. There is also, however, a conceptual reason to rule out credences that assign zero probability to the actual evidence. It seems that actually observing something that you initially took to be impossible¹⁶ might give you reason to reassess whether you had sensible prior beliefs. Surely it would have been better to give *some* credence to the event that actually occurred? This motivates the idea that prior probability functions ought to be *regular*, that is, they ought to assign zero probability to logical falsehoods only. Now, regularity is commonly known to be an impossible property to satisfy when your space of events is infinitely large, but closer inspection reveals that only a *restricted* (and satisfiable) form of regularity is motivated by the above. Any evidence proposition you might actually observe is a member of some finite partition over which you ought to have non-zero probabilities. I don't know about you, but I have never observed an infinitely thin dart being thrown at the unit interval; I have never bought a ticket to a lottery with countably many tickets; I've never observed all of an infinite sequence of coin tosses. Even in cases where the full space of possibilities is infinite (for example, arbitrarily long sequences of coin flips) I only ever actually observe some element of a finite partition of the evidence (some initial segment of an infinite sequence for example). We have to be somewhat careful here. It is trivial that any proposition A is a member of some finite partition ($\{A, \neg A\}$ for instance), but what I am proposing here is that every evidence proposition we care about is a member of a finite partition *such*

15 Even if you adopt some theory of conditional probability that sidesteps the zero probability issue – primitive conditional probabilities; taking $\mathbf{pr}(X|E)\mathbf{pr}(E) = \mathbf{pr}(XE)$ as definitional – you're still often left with the issue that conditioning on an event of zero probability is seriously under-constrained by the probability axioms.

16 The relationship between impossibility and zero probability is not straightforward, particularly in the context of sufficiently big spaces of events, but that is not the issue I am pointing to here. See Hájek (2003, n.d.) and Lyon (2014).

that it is sensible to require that your prior probability in every partition element is non-zero. For example, any proposition of the form “observe h heads in n tosses of the coin” is a proposition that has this property: the finite partition it belongs to is the $(n + 1)$ -membered partition of all possible numbers of heads in n tosses. And if we focus our attention on your (conditional) credences about other possible evidence propositions (other members of finite evidence partitions) then it is perfectly reasonable and possible to require all of them to be non-zero. Note that all the learning priors from the previous section are regular in this restricted sense.

A note on terminology. If we have observed evidence E , it is standard to call the function that assigns to each probability function $\mathbf{pr}$ the value $\mathbf{pr}(E)$, the “likelihood” function. This function is very commonly used as a method of determining how “good” a probability function is, given some evidence. Sometimes this is the basis of the whole approach to statistics (e.g. Pawitan 2001), and even when it is not, the likelihood is an important part of the statistical toolkit. So the remarks of the previous paragraph could be rephrased like so: there’s a conceptual reason to consider probability functions that have zero likelihood to be defective.

Recall that we defined “Generalised Conditioning” (GC) as:

$$\mathbb{P}(X|E) = \{\mathbf{pr}(X|E), \mathbf{pr} \in \mathbb{P}, \mathbf{pr}(E) > 0\}$$

You will note that last condition builds in the caveat we just discussed: $\mathbf{pr}(E) > 0$. But note that such an update rule is one of many taking a simple form:

$$\mathbb{P}(X|_{\mathbb{Q}}E) = \{\mathbf{pr}(X|E), \mathbf{pr} \in \mathbb{Q} \subseteq \mathbb{P}\}$$

We recover generalised conditioning by setting $\mathbb{Q} = \{\mathbf{pr} \in \mathbb{P}, \mathbf{pr}(E) > 0\}$. Note that such a $\mathbb{Q}$ depends on both E and $\mathbb{P}$. We might consider exploring alternatives to GC by considering alternative choices of $\mathbb{Q}$. Some alternatives are discussed by, for example, Bradley and Steele (2014b), Gilboa and Schmeidler (1993), and Grove and Halpern (1998). Here are four candidates for $\mathbb{Q}$.

1. $\mathbb{Q} = \{\mathbf{pr} \in \mathbb{P}, \mathbf{pr}(E) > 0\}$
2. $\mathbb{Q} = \{\mathbf{pr} \in \mathbb{P}, \mathbf{pr}(E) > \tau\}$ for some fixed threshold $0 \leq \tau \leq 1$.
3. $\mathbb{Q} = \{\mathbf{pr} \in \mathbb{P}, \mathbf{pr}(E) = \overline{\mathbb{P}}(E), \mathbf{pr}(E) > 0\}$
4. $\mathbb{Q} = \{\mathbf{pr} \in \mathbb{P}, \mathbf{pr}(E) \geq \alpha \overline{\mathbb{P}}(E), \mathbf{pr}(E) > 0\}$, for $\alpha \in (0, 1)$

The first of these is Generalised Conditioning. While the second seems to get something right – priors should pass some test of “reasonableness” as determined by the evidence received – it has the disadvantage of sometimes being empty (when $\overline{\mathbb{P}}(E) \leq \tau$). The third is the “Maximum likelihood” rule discussed by Gilboa and Schmeidler (1993). The name makes sense, since the rule proposes that you should update your credal set to the set of probability functions that attain the maximum likelihood value among the admissible priors. It yields an empty set in

fewer circumstances,¹⁷ but it is perhaps too hasty in its jumping to conclusions. The last option is what Cattaneo (2014) calls “ α -cut conditioning”, and we shall denote it $\mathbb{P}(X|_{\alpha}E)$. This update rule is the one I want to focus on. Note that if you make α very small, this will be very similar to generalised conditioning. And if you make α very close to one, then α -cut conditioning will be very close to maximum likelihood.¹⁸ The α parameter could be thought of as encapsulating a kind of epistemic “conservativity” (although not the same kind of conservativity that Konek n.d. discusses). The bigger your α , the quicker you jump to conclusions.

Several authors have suggested equipping a set of probabilities with some kind of measure of reliability Bradley (2017); Gärdenfors and Sahlin (1982); Hill (2013), and Lyon (2017). This proposal is interesting, as far as it goes, but these authors don’t really provide much explanation of where this index of reliability is supposed to come from. As Cattaneo (2008) points out, the likelihood $\mathbf{pr}(E)$ gives you a measure of how reliable each probability in $\mathbb{P}$ is given the actual evidence. Thus, α -cut conditioning can be seen as part of a likelihood-based approach to enriching sets of probabilities with a well-defined and objective measure of reliability. Hill (n.d.) proposes an updating procedure which is similar to α -cut conditioning, but somewhat more complicated since the basic object is a nested set of sets of probabilities, rather than a single set of probabilities.

One can motivate α -cut conditioning in a very similar way to the more conceptual motivation for ruling out zero probability conditioning in the precise case. Those $\mathbf{pr} \in \mathbb{P}$ that gave relatively small probability to E seem somewhat defective, and it is reasonable to excise them from your credal state. How small the prior probability of the evidence has to be is determined by the α parameter, and $\mathbb{P}(E)$ which is a kind of normalising factor. This normalising factor is necessary since, as your evidence gets more informative, even the most likely possibilities become less likely in absolute terms. For example, imagine that you know the chance of the coin landing heads is 0.3. The probability you assign to the evidence of one head in one toss is 0.3. This is the less likely of the possibilities (one tail in one toss has probability 0.7). But consider the evidence of five heads in 16 tosses. This is the most likely outcome of 16 tosses of a coin with chance 0.3 of landing heads, and yet, the probability of such an outcome is around 0.21: so less likely (in absolute terms) than one head in one toss. Note that this normalising process means that the α -cut conditioning is not a pointwise constraint in the terms of Moss (n.d.).

¹⁷ If $\mathbb{P}(E) = 0$ or if $\mathbb{P}(E)$ does not attain its upper bound, this $\mathbb{Q}$ will be empty. One might fix one of these problems by taking the closure of $\mathbb{P}(E)$.

¹⁸ This rule has the advantage that it yields an empty set only when $\mathbb{P}(E) = 0$.

Sceptics of the value of regularity (even in this restricted form) can still grant that there is something plausible about considering probabilities in the credal set with extremely low relative likelihood to be defective.

We can also see the plausibility of this method of updating by reflecting on the “credal committee” metaphor (Joyce 2010). Imagine if some committee member was such that they considered what actually happened to be significantly more unlikely than the other members of the committee did. This seems to be a kind of epistemic failing, and one that might lead to that committee member being ignored or thrown off the committee. The α parameter governs how widely you consult among your credal committee. A high α means you only listen to those who were close to being the most reliable, and thus you tend to jump to conclusions, whereas a lower α means you pay attention to more of your committee, even if they were not as reliable; this builds in a kind of conservatism.

One might want to pursue the option of more sophisticated methods for restricting the set of conditionalised probabilities, for example by using some form of accuracy-based scoring rule approach (Pettigrew 2016); or by letting α depend on E . Given our understanding of α as a measure of epistemic conservatism, it only makes sense to consider varying α when we think it is sensible to allow varying rates of conservatism within the same agent. α might vary by subject matter, for instance. I won’t pursue this thought further here because it seems plausible to me that an agent ought to have similar levels of conservatism to all propositions about coin tosses that come up in the examples we discuss.

Let’s return to the coin tossing example from the earlier section. We’re going to show that α -cut conditioning does appear to learn in simulations. Before we do that, a brief word on iterating α -cut conditionalisation. What happens if you update by E and then update by F ? Does that give the same result as learning F and then E ? It depends on how you update. Let $\mathbb{P}'(-) = \mathbb{P}(-|_{\alpha}E)$. When you then learn F , what do you do? $\mathbb{P}(-|_{\alpha}EF)$ or $\mathbb{P}'(-|_{\alpha}F)$? Call the former of these “total evidence” update, and the latter “iterative” update. In what follows, we’ll focus on total evidence update; for total evidence update, the order you learn propositions does not matter. See Section 5.2 for more on iterative update.

Analytic results for our coin tossing example are hard to come by, but if we simulate α -cut conditioning, we see that it converges (Figure 3).¹⁹ The grey lines represent what generalised conditioning looks like for the probabilities in the set,

¹⁹ Each red line represents a probability function that has not yet been removed by α -cut. Each grey line represents a probability that has been removed. The blue lines represent two probabilities that start out at the extremes for $\mathbf{pr}(H)$, and that converge.

Figure 3: Alpha cut learning.

and the red lines are those that make it through the α -cut process.²⁰ So we can see that α -cut conditioning does appear to converge on the true value of heads. If we graph the spread of probability values for heads over time (Figure 4), we see that it appears to shrink at a rate of about $\frac{1}{\sqrt{n}}$ for total evidence update (see also Wilson 2001). The fact that the spread for GC appears to shrink at all is an artefact of only simulating finitely many priors: there are more stubborn priors, not simulated, that prevent GC from shrinking at all. Figure 5 demonstrates that this is not a one off: on average, the spread $\mathbb{P}(H|S_n) - \underline{\mathbb{P}}(H|S_n)$ shrinks as evidence grows, and the interval $[\underline{\mathbb{P}}(H|S_n), \mathbb{P}(H|S_n)]$ almost always covers the true chance $\mathbf{pr}^*(H)$.²¹

20 There should be no gaps to the GC grey lines: for every amount of evidence and every value in $[0, 1]$ there should be a prior whose GC update takes that value after that evidence. The gaps are due to the fact that I can only simulate finitely many priors. The blue lines represent the upper and lower probabilities for the set of all learning priors with $\lambda < 8$. Setting a maximum value for λ is an alternative to the current approach that I explore in [BLINDED].

21 This graph shows the average value of spread and coverage in a number of trials, with varying values of the true chance. The vertical lines represent a range in which 95% of the observed values fall. In this figure each number along the x axis represents learning the outcome of eight tosses of the coin, but the actual values are random, drawn from a binomial distribution with parameter $\mathbf{pr}^*$.

Figure 4: $\bar{\mathbb{P}}(H) - \mathbb{P}(H)$ as evidence accrues.

So our question becomes “under what circumstances does α -cut allow you to learn H on the basis of evidence?” I don’t have any firm results here, but here are some gestures toward future work. If your credal set has some kind of natural “metric” on it such that it makes sense to talk about how close one function is to another (for example if they’re all parameterized by some compact subset of $\mathbb{R}^n$) then call a likelihood function “peaked around $\mathbf{pr}$ ” if only those probabilities close to $\mathbf{pr}$ have high relative likelihood. In circumstances amenable to learning, you are likely to receive evidence such that the likelihood is peaked around a probability function near to the truth $\mathbf{pr}^*$.²² If $\mathbf{pr}^*(H) = r$, then $\frac{S_n}{n}$ is likely to be near r because of the law of large numbers, for example. If it is further the case that $|\mathbf{pr}(H|E_n) - \mathbf{pr}^*(H)|$ is small when $\mathbf{pr}$ is close to $\mathbf{pr}^*$, then $\mathbb{P}$ will exhibit sensible learning behaviour. Obviously, much work will need to be done to tighten up what kind of “closeness” is appropriate for which learning problems, and how strong the

(H). “Coverage” here is defined as the maximum of 0, $\mathbf{pr}(H) - \bar{\mathbb{P}}(H|S_n)$ and $\mathbb{P}(H|S_n) - \mathbf{pr}^*(H)$: that is, if it is 0, the true chance is between the upper and lower probability.

²² This property is very closely connected to the standard statistical concept of the maximum likelihood estimator’s being consistent.

Figure 5: Average improvement for alpha cut.

bounds can be made. I take it that the above gestures towards a formal result at least suggest that this approach is worth exploring further.

4 Coherence

What coherence property does IP with α -cut conditioning satisfy? In general terms, the answer is that α -cut conditioning doesn't even satisfy Avoid Sure Loss, as the following example demonstrates. Lottery A is a fair lottery with n tickets. You know nothing about lottery B, except that it also has n tickets so you have a near-ignorance prior for outcomes from lottery B. I'm going to flip a fair coin to decide whether to draw a lottery A ticket (heads) or a lottery B ticket (tails). I'm also going to offer you bets denominated in some units such that your utility is linear with respect to those units. Before I flip the coin, I'll offer you a bet at even odds that the coin lands heads; i.e. you pay $\frac{1}{2}$ unit to win one unit if the coin lands heads. Before revealing the outcome of the coin toss, I'm then going to tell you the outcome of the draw from the urn. I then offer you a bet against heads at worse than even odds. (Pay $1 - \frac{1}{2n\alpha}$ units to win one unit if tails). Unless α is small enough (i.e. less than $1/n$), these bets lose you money, whichever outcome I announce.²³ Why does this pathological behaviour manifest? What's going on is that once you learn that ticket

²³ Thanks to Marco Cattaneo for this example.

i was drawn from some lottery, the highest likelihood probabilities are those that take lottery B to be extremely biased towards i . This plus the fact that the chance set up makes the announcement of the lottery outcome *dilate* the coin toss probabilities (Pedersen and Wheeler 2014; Seidenfeld and Wasserman 1993), leads to some weird behaviour.

In our mystery coin example, however, α -cut conditioning does Avoid Sure Loss. According to Zaffalon and Miranda (2013, Theorem 1), you avoid sure loss if and only if your unconditional upper prevision for every f is greater than the lower prevision you would have for f if you were to learn B for at least some B . α -cut conditioning satisfies that condition in the mystery coin example, but not in general. So at least for sets of learning priors and coin-toss evidence, α -cut conditioning Avoids Sure Loss. If you know you'll learn one element of some partition $\mathcal{B}$, and you know you'll be offered the chance to accept or reject some finite menu of gambles, then it's always possible to set α so as to not be subject to sure loss. The intuition here is that given that as α gets smaller, your posterior credal set more closely resembles the outcome of GC, at least for finite partitions.²⁴ This solution is theoretically unappealing though, since it requires setting α to some particular value so as to avoid sure loss, whereas α was supposed to be a parameter determined by your taste for epistemic risk.

However, α -cut conditioning cannot satisfy coherence, unless it coincides with GC. Generalised Conditioning is the most informative possible jointly coherent conditional prevision (Miranda 2009, Lemma 2). What this means is that any conditional lower probability that is jointly coherent with prior $\mathbb{P}$, must be such that its lower conditional probabilities are lower than those that GC yields. Since the whole point of α -cut conditioning was to provide *more informative* conditional beliefs than those obtained by GC, we cannot require that α -cut satisfy this stronger condition.

So α -cut conditionalisers will sometimes be such that they don't find a set of bets that guarantees a net gain to be acceptable. To some extent, this is an issue that imprecise probabilists will have to address anyway, since a similar kind of apparent problem arises for decisions made over time, even without any updating (Elga 2010): in both cases the issue is that combining judgements of acceptability across time yield unintuitive results. This sort of sequential choice incoherence – or “dynamic inconsistency” – is discussed in a number of places. In philosophy, several people have responded directly to Elga's challenge (Moss 2015a; Rinard 2015, for example); in economics and psychology the idea is often tied up with

²⁴ For infinite partitions, even GC is sometimes not jointly coherent with the prior (Walley 1991, Appendix J) so this case will require a lot more care, or perhaps a move to a more general framework such as sets of desirable gambles.

discussions of “ambiguity aversion” (Al-Najjar and Weinstein 2009; Trautmann and Van der Kijlen 2016). A full discussion of this topic is beyond the scope of the current paper, but note that the advocate of α -cut conditionalisation has many resources to bring to bear on the issue of dynamic inconsistency. One approach – that of Bradley and Steele (2014a) and Rinard (2015), for example – is to deny the “package principle” that governs how the acceptability of a series of bets is determined by the acceptability of its members.

Whether α -cut conditioning avoids sure loss depends on the kind of propositions one might learn. It is an open question on whether it’s always possible to set α low enough to avoid sure loss in a given learning scenario, so exactly what weakened kind of coherence we can salvage is currently unknown. It would also be interesting to systematically explore *how* incoherent α -cut conditioning is in the sense of Schervish, Seidenfeld, and Kadane (1997) and Staffel (2020). A natural hypothesis is that the smaller α is, the less incoherent α -cut conditioning is.

5 Conclusions

My conclusions are modest. For the case of learning the bias of a coin from evidence of repeated tosses, α -cut conditioning appears to converge on the truth.²⁵ This does mean, however, giving up on coherence and making do with just avoiding sure loss. This should provide some small amount of comfort to the imprecise probabilist: belief inertia is not an insurmountable challenge. Beyond this simple case of coin tossing, results are more mixed. I suspect there will be a more general convergence result, but without more clarity on when this approach to update avoids sure loss in general, such a learning result is of limited interest. I believe that this approach to update should be explored in more detail in the future.

Appendix A: Merging of Probabilities

Schervish and Seidenfeld (1990) offer some results that might appear to make belief inertia a phenomenon that can only occur under relatively circumscribed conditions. We first state a result that follows from Schervish and Seidenfeld (1990), Corollary 1.

²⁵ More generally, we can expect the same conclusions to hold for learning multinomial data with sets of Dirichlet priors. While this sounds like quite a restricted result, priors of this kind have extremely wide applicability (Walley 1996), and so this result is fairly general in practice.

Merging of Opinion

Your prices for gambles in a one-sided betting scenario should be such that no combination of acceptable bets is guaranteed to yield a loss in every state.

If $\mathbb{P}$ is contained in a closed, convex set of (countably additive) probabilities $\mathbb{Q}$ such that:

- for all $\mathbf{pr}, \mathbf{pr}' \in \mathbb{Q}$ we have $\mathbf{pr}(X) = 0$ if $\mathbf{pr}'(X) = 0$ for all X in $\mathcal{A}$ (i.e. all the elements of $\mathbb{Q}$ are mutually absolutely continuous)
- $\mathbb{Q}$ is the convex hull of some finite subset of its elements (i.e. $\mathbb{Q}$ has finitely many extreme points)

then for almost all sequences of possible evidence, the probabilities in $\mathbb{P}$ converge in the sense that for any $\mathbf{pr}, \mathbf{pr}' \in \mathbb{P}$, any $X \in \mathcal{A}$:

$$|\mathbf{pr}(X|E_n) - \mathbf{pr}'(X|E_n)| \rightarrow 0$$

That is, if the probabilities agree on the measure zero events, then, with increasing evidence, they will get closer together on the probability for every event. So merging guarantees narrowing, but not necessarily convergence. This merging of opinion only constitutes learning if, among the $\mathbf{pr} \in \mathbb{P}$, there is some probability such that being close to it is epistemically good. We earlier called this the “expert probability”. This result secures that probabilities meeting the above conditions will merge (i.e. get closer together) and, if the expert probability is in the set, then we can call this merging “learning”. “Belief inertia” is, then, a failure to merge even in the presence of such expert probabilities. So, to the extent that it’s reasonable to expect that (a) your prior credal set meet the above conditions, and that (b) your prior credal set contain suitable expert probabilities, the belief inertia is no problem.

As we saw, belief inertia does happen for the learning priors discussed in the main text. Which of (a) or (b) is false in this example? Let’s start with (a): the conditions for the application of the Schervish and Seidenfeld (1990) result don’t hold. Consider any two probability functions $\mathbf{pr}, \mathbf{q}$, that make the coin tosses independent and identically distributed. That is, these probability functions are candidates to be the chance function. If you like, functions of this kind are the possible infinite limits of the process of learning the outcomes of coin tosses.²⁶ They are infinitely stubborn in the sense that no amount of evidence will change their probability for the next toss landing heads.

²⁶ If only these i.i.d. probabilities are candidate expert probabilities, then (b) is thus false. But there are credences in the set that could reasonably be called experts: those that have γ close to $\mathbf{pr}^*(H)$ for example.

No two distinct such probability functions are mutually absolutely continuous since if $\mathbf{pr}(H_i) = p \neq q = \mathbf{q}(H_i)$, then $\mathbf{pr}$ gives probability one to the set of sequences of coin tosses with limiting frequency p , and $\mathbf{q}$ gives that set probability zero (and vice versa). So, even though the set of beta priors are all in the convex hull of the set of these chance distributions (this is a consequence of de Finetti's representation theorem, since coin tosses are exchangeable) these extrema are not mutually absolutely continuous, and thus the above result does not apply.

Beyond the realm of learning i.i.d. evidence, the above result highlights circumstances under which narrowing is guaranteed. If you are sure that the expert probability is among those with which you are merging, then successful inductive inference is assured. I have no answer to the question of how often these conditions are satisfied in real examples.

Appendix B: Commutativity

We are going to need a few definitions, so I collect them here:

- $\mathbb{Q}_\alpha[E, \mathbb{P}] = \{\mathbf{pr} \in \mathbb{P}, \mathbf{pr}(E) \geq \alpha \bar{\mathbb{P}}(E)\}$
- $\mathbb{P}_E^\alpha = \mathbb{P}(-|_\alpha E) = \{\mathbf{pr}(-|E), \mathbf{pr} \in \mathbb{P}, \mathbf{pr}(E) \geq \alpha \bar{\mathbb{P}}(E)\}$

Consider the case of learning E , then learning F . In Figure 6, iterative conditioning amounts to moving diagonally down and left to $\mathbb{P}_E^\alpha$ and then down and left again to $\mathbb{P}_E^\alpha(-|_\alpha F)$. Total evidence conditioning amounts to moving diagonally down and left to $\mathbb{P}_E^\alpha$, but then learning F amounts to α -cut conditioning the prior $\mathbb{P}$ on total

Figure 6: α -cut conditioning as GC plus restriction.

I don't consider this failure of commutativity to be an insurmountable problem since, if the priors in the credal set are suitably well-behaved (i.e. satisfy the conditions for some form of convergence theorem) then, whatever order the evidence comes in, the outcome will be a set of probability functions each getting closer to the truth. However, coupled with the fact that iterative update doesn't necessarily converge on the truth, I think we have reason to prefer total evidence update. Given that total evidence update appears to be what time-slice epistemologists like Moss (2015b) and Hedden (2015), advocate as the right way to conceive of update, one might think that this serves as an argument in favour of their view. I don't pursue this line of thought since I think there are good reasons to think there really are diachronic norms for decision making, but that discussion is beyond the scope of this paper.

Let's now consider whether iterative α -cut conditioning converges. If you end up in a situation where $\alpha\overline{\mathbb{P}}(E) \leq \mathbb{P}(E)$ then nothing is thrown out. Consider updating (iteratively) after every toss of a coin. After a few tosses, you've thrown out the extreme priors, but there's still a range of priors available. So your priors shrink for a bit, but then stop shrinking. You can see this happening in Figure 4: the line for iterative update starts dropping as fast as the total evidence spread does, but after a certain time, it levels off and starts to drop at about the rate GC does. This doesn't happen in the total evidence updating case because the "amount of evidence" you update on keeps growing. As Figure 8 illustrates iterative update

Figure 8: Iterative and total evidence update compared.

eventually stops converging much, whereas total evidence update continues to converge. (Indeed, that iterative update appears to converge at all in the right-hand portion of the graph is due to the artefact that only a finite number of priors are simulated).

References

- Augustin, T., F. P. A. Coolen, G. de Cooman, and M. C. M. Troffaes, eds. 2014. *Introduction to Imprecise Probabilities*. New York: John Wiley and Sons.
- Al-Najjar, N. I., and J. Weinstein. 2009. "The Ambiguity Aversion Literature: A Critical Assessment." *Economics and Philosophy* 25: 249–84.
- Bradley, R. 2017. *Decision Theory with a Human Face*. Oxford: Oxford University Press.
- Bradley, S. 2015. "How to Choose Among Choice Functions." In *ISIPTA 2015 Proceedings*, 57–66.
- Bradley, S. 2019. "Imprecise Probabilities." In *The Stanford Encyclopedia of Philosophy*, edited by E. N. Zalta. The Metaphysics Research Lab.
- Bradley, S., and K. Steele. 2014a. "Should Subjective Probabilities Be Sharp?" *Episteme* 11: 277–89.
- Bradley, S., and K. Steele. 2014b. "Uncertainty, Learning and the 'Problem' of Dilation." *Erkenntnis* 79: 1287–303.
- Buchak, L. 2013. *Risk and Rationality*. Oxford: Oxford University Press.
- Castro, C., and C. Hart. 2019. "The Imprecise Impermissivist's Dilemma." *Synthese* 196 (4): 1623–40.
- Cattaneo, M. 2008. "Fuzzy Probabilities Based on the Likelihood Function." In *Soft Methods for Handling Variability and Imprecision*, edited by D. Dubois, M. A. Lubiano, H. Prade, M. Á. Gil, P. Grzegorzewski, and O. Hryniewicz, 43–50. Berlin, Heidelberg: Springer.
- Cattaneo, M. 2014. "A Continuous Updating Rule for Imprecise Probabilities." In *Information Processing and Management of Uncertainty in Knowledge Based Systems*, edited by A. Laurent, O. Strauss, B. Bouchon-Meunier, and R. R. Yager, et al., 426–35. Springer.
- Elga, A. 2010. "Subjective Probabilities Should Be Sharp." *Philosophers' Imprint* 10: 1–11.
- Gärdenfors, P., and N.-E. Sahlin. 1982. "Unreliable Probabilities, Risk Taking and Decision Making." *Synthese* 53: 361–86.
- Gilboa, I., and D. Schmeidler. 1993. "Updating Ambiguous Beliefs." *Journal of Economic Theory* 59: 33–49.
- Grove, A., and J. Y. Halpern. 1998. "Updating Sets of Probabilities." In *Proceedings of the fourteenth conference on uncertainty in AI*, 173–82.
- Hájek, A. 2003. "What Conditional Probabilities Could Not Be." *Synthese* 137: 273–323.
- Hájek, A. 2008. "Arguments for—or Against—Probabilism?" *The British Journal for the Philosophy of Science* 59: 793–819.
- Hájek, A. n.d. Staying Regular?
- Hedden, B. 2015. *Reasons without Persons*. Oxford: Oxford University Press.
- Hill, B. n.d. *Updating Confidence in Beliefs*. Also available at <https://ssrn.com/abstract=3385116>.
- Hill, B. 2013. "Confidence and Decision." *Games and Economic Behavior* 82: 675–92.
- Joyce, J. M. 2005. "How Probabilities Reflect Evidence." *Philosophical Perspectives* 19: 153–78.

- Joyce, J. M. 2010. "A Defense of Imprecise Credence in Inference and Decision." *Philosophical Perspectives* 24: 281–323.
- Konek, J. n.d. "Epistemic Conservativity and Imprecise Credence." *Philosophy and Phenomenological Research*.
- Lassiter, D. 2020. "Representing Credal Imprecision: from Sets of Measures to Hierarchical Bayesian Models." *Philosophical Studies* 177: 1463–85.
- Levi, I. 1974. "On Indeterminate Probabilities." *Journal of Philosophy* 71: 391–418.
- Levi, I. 1980. *The Enterprise of Knowledge*. Cambridge: The MIT Press.
- Lyon, A. 2014. "From Kolmogorov to Popper to Rényi: There's No Escaping Humphrey's Paradox (When Generalized)." In *Chance and Temporal Asymmetry*, edited by T. Handfield, and A. Wilson. Oxford: Oxford University Press, 112–25.
- Lyon, A. 2017. "Vague Credence." *Synthese* 194 (10): 3931–54.
- Miranda, E. 2009. "Updating Coherent Previsions on Finite Spaces." *Fuzzy Sets and Systems* 160: 1286–307.
- Moss, S. 2015a. "Credal Dilemmas." *Noûs* 49: 665–83.
- Moss, S. 2015b. "Time-Slice Epistemology and Action Under Indeterminacy." In *Oxford Studies in Epistemology*, 172–94.
- Moss, S. n.d. "Global Constraints on Imprecise Credences: Solving Reflection Violations, Belief Inertia and Other Principles." *Philosophy and Phenomenological Research*.
- Pawitan, Y. 2001. *All Likelihood: Statistical Modelling and Inference Using Likelihood*. Clarendon Press.
- Pedersen, A. P., and G. Wheeler. 2014. "Demystifying Dilation." *Erkenntnis* 79: 1305–42.
- Pettigrew, R. 2016. *Accuracy and the Laws of Credence*. Oxford: Oxford University Press.
- Rinard, S. 2013. "Against Radical Credal Imprecision." *Thought* 2: 157–65.
- Rinard, S. 2015. "A Decision Theory for Imprecise Probabilities." *Philosophers' Imprint* 15: 1–16.
- Schervish, M. J., and T. Seidenfeld. 1990. "An Approach to Consensus and Certainty with Increasing Evidence." *Journal of Statistical Planning and Inference* 25: 401–14.
- Schervish, M. J., T. Seidenfeld, and J. B. Kadane. 1997. *Two Measures of Incoherence: How Not to Gamble if You Must, Technical Report*. Pittsburgh: Department of Statistics, Carnegie Mellon University.
- Seidenfeld, T., and L. Wasserman. 1993. "Dilation for Sets of Probabilities." *Annals of Statistics* 21: 1139–54.
- Staffel, J. 2020. *Unsettled Thoughts*. Oxford: Oxford University Press.
- Trautmann, S., and G. Van der Kuylen. 2016. "Ambiguity Attitudes." In *Blackwell Handbook of Judgement and Decision-Making*, edited by G. Keren, and G. Wu, 89–116. Blackwell.
- Troffaes, M., and G. de Cooman. 2014. *Lower Previsions*. Wiley.
- Vallinder, A. 2018. "Imprecise Bayesianism and Global Belief Inertia." *The British Journal for the Philosophy of Science* 69 (4): 1205–30.
- Walley, P. 1991. "Statistical Reasoning with Imprecise Probabilities." In *Monographs on Statistics and Applied Probability*, Vol. 42. Chapman and Hall.
- Walley, P. 1996. "Inferences from Multinomial Data: Learning about a Bag of Marbles." *Journal of the Royal Statistical Society B* 58 (1): 3–57.
- Wilson, N. 2001. "Modified Upper and Lower Probabilities Based on Imprecise Likelihoods." In *Proceedings of the 2nd International Symposium on Imprecise Probabilities and their Applications*.
- Zaffalon, M., and E. Miranda. 2013. "Probability and Time." *Artificial Intelligence* 198: 1–51.