“Weight of Evidence, Evidential Completeness and Accuracy”

Rafal Urbaniak and Marcello Di Bello

1 Motivations

1.1 Balance vs. weight

Suppose we want to represent our uncertainty about a proposition in terms of a single probability that we assign to it. It is not too difficult to inspire the intuition that this representation does not capture an important dimension of how our uncertainty connects with the evidence we have or have not obtained. In a 1872 manuscript of *The Fixation of Belief* (W3 295) C. S. Peirce gives an example meant to do exactly that.

> When we have drawn a thousand times, if about half have been white, we have great confidence in this result. We now feel pretty sure that, if we were to make a large number of bets upon the color of single beans drawn from the bag, we could approximately insure ourselves in the long run by betting each time upon the white, a confidence which would be entirely wanting if, instead of sampling the bag by 1000 drawings, we had done so by only two.

The objection is not too complicated. Your best estimate of the probability of $W =$‘the next bean will be white’ is .5 if half of the beans you have drawn randomly so far have been white, no matter whether you have drawn a thousand or only two of them. But this means that expressing your uncertainty about W by locutions such as “my confidence in W is .5’ does not capture this intuitively important distinction.

Similar remarks can be found in Peirce’s 1878 *Probability of Induction*. There, he also proposes to represent uncertainty by at least two numbers, the first depending on the inferred probability, and the second measuring the amount of knowledge obtained; as the latter, Peirce proposed to use some dispersion-related measure of error (but then suggested that an error of that estimate should also be estimated and so, so that ideally more numbers representing errors would be needed).

Peirce himself did not call this the weight of evidence (and in fact, used the phrase rather to refer to the balance of evidence, W3 294) [CITE KASSER 2015]. However, his criticism of such an oversimplified representation of uncertainty anticipated what came to be called weight of evidence by Keynes in his 1921 *A Treatise on Probability*:

> As the relevant evidence at our disposal increases, the magnitude of the probability of the argument may either increase or decrease, according as the new knowledge strengthens the unfavourable or the favourable evidence; but something seems to have increased in either case,—we have a more substantial basis upon which to rest our conclusion. I express this by saying that an accession of new evidence increases the weight of an argument. New evidence will sometimes decrease the probability of an argument but it will always increase its ‘weight.’ (p. 71)

The key point is the same [CITE LEVI 2001]: the balance of probability alone cannot characterize all important aspects of evidential appraisal. Keynes also considered measuring weight of evidence in terms of the variance of the posterior distribution of a certain parameter, but was quite attached to the idea that weight should increase with new information, even if the dispersion increase with new evidence [TP 80-82], and so he proposed only a very rough sketch of a positive sketch. Moreover, as he was uncertain how a measure of weight should be incorporated in further decision-making, the was skeptical about the practical significance of the notion. [TP 83]

But what is this positive sketch? On one hand, Keynes [TP 58-59] connects the notion of weight with relevance. Call evidence E relevant to X given K just in case $\mathsf{Pr}(X|K \wedge E) \neq \mathsf{Pr}(X|K)$.[1] One postulate than can be found in the *Treatise* [TP 84] is:[2]

(Monotonicity) If E is relevant to X given K, where K is background knowledge, $V(X|K \wedge E) > V(X|K)$, where V is the weight of evidence.

[RUNDE 1990, 280] suggests that Keynes at some point calls weight the completeness of information. This however, is a bit hasty, as Keynes only says that *the degree of completeness of the information on which a probability is based does seem to be relevant, as well as the actual magnitude of the probabiltiy, in making practical decisions*. As later on we will argue that it is actually useful to distinguish evidential weight (how much evidence do we have?) and evidential completeness (do we have all the evidence that we would expect in a given case?), we rather prefer to extract a more modest postulate:

(Completeness) If E_1 and E_2 are relevant items of evidence, and E_2 is (in a sense to be discussed) more complete than E_1, $V(X|K \wedge E_2) > V(X|K \wedge E_1)$.

If we conceptualize E_2 being complete and E_1 being incomplete as E_2 being a maximal relevant conjunction of relevant claims one of which is E_1, (Completeness) follows from (Monotonicity).

Similar requirements seem to be inspired by the urn example. We put them in two forms, a weaker and a stronger one.

(Weak increase) In cases analogous to the urn example, the weight obtained by a larger sample is higher, if the frequencies in the samples remain the same.

(Strong increase) In cases analogous to the urn example, the weight obtained by a larger sample is higher.

Now, some requirements on how weight of evidence is related to the balance of probability. For one thing, Keynes insists that new (relevant) evidence might decrease probability but will always increase weight [TP 77]. Since (Monotonicity) already captures the idea that weight will always increase, here we extract the other part of the claim:

(Possible decrease) It is possible that $V(X|K \wedge E) > V(X|K)$ while $\mathsf{P}(X|K \wedge E) < \mathsf{P}(X|K)$.

Clearly, Keynes also endorsed the following two requirements of a very similar form:

(Possible increase) It is possible that $V(X|K \wedge E) > V(X|K)$ while $\mathsf{P}(X|K \wedge E) > \mathsf{P}(X|K)$.

(Possibly no change) It is possible that $V(X|K \wedge E) > V(X|K)$ while $\mathsf{P}(X|K \wedge E) = \mathsf{P}(X|K)$.

Interestingly, Keynes for quite a few years did not attempt to provide anything close to a formal explication of the notion, and did not spend too much time studying the issue. Various reasons for this has been proposed the literature, a prominent one [CITE FEDUZI 2010] being that from the decision-theoretic perspective no clear stopping rule emerged as to whether the evidence is weighty enough to make a decision. Later on we will see a sort of revival—some ideas later developed by Keynes has been used to explicate the notion of weight formally, and we will take a closer look at this proposal.

Should I talk about other theories here, or should we leave it as is without getting into interpretative details.

REF section

1.2 Examples and informal desiderata

- Go over Nance in particular, some other sources?
- first check for completeness, then evaluate
- what do you mean: are there items of relevant evidence that you could reasonably obtain
- destroyed?

1.2.1 Monotonicity of weight

Before we move on, let us ponder whether (Monotonicity) is actually desirable. Is it always the case, as some formulations from Keynes would suggest, that any item of relevant of evidence, when obtained,

[1] Keynes also uses a slightly more convoluted notion of relevance to avoid equally strong items of opposite evidence turning out to be irrelevant (this objection has also been brought up by [COHEN 1986 TWELVE]). The more complex version is that a proposition E_1 is relevant to X given K just in case it entails a proposition E_2 such that $\mathsf{P}(X|K \wedge E_2) \neq \mathsf{Pr}(X|K)$. [COHEN 1986 TWELVE] complaints that this still runs into difficulties. Ignore K, take an irrelevant proposition Z. It entails $Z \vee X$ and $\mathsf{P}(Z \vee X|X \wedge E) = 1$. Now, by Bayes' theorem we have $\mathsf{P}(X|E \wedge (Z \vee X)) = \frac{\mathsf{P}(X|E) \times \mathsf{P}(Z \vee X|X \wedge E)}{\mathsf{P}(Z \vee X|E)} = \frac{\mathsf{P}(X|E)}{\mathsf{P}(Z \vee X|E)}$. If the denominator differs from 1, the result differs from the numerator. We will ignore such difficulties, as they are not of key importance for the development of this chapter.

[2] RUNDE 1990 283 suggests Keynes allows for weight of evidence to decrease when new evidence increases the range of alternatives, but this is based on Keynes' claim that weight is increased when the number of alternatives is reduced, and Keynes does not directly say anything about the possibility of an increase of the number of alternatives.

leads to a higher weight of evidence?

Here are two examples from [WEATHERSON 2002], one qualitative, one quantitative. First. you are playing poker and wonder if one of the other players, Lydia, has a straight flush (five cards of sequential rank in the same suite). There are 40 possible straight flush hands out of 2,958,960 possible hands, so you estimate the probability of this event to be $40/2{,}958{,}960$. But then you look at her facial expressions, listen to her tone of voice, past bluffing behavior, and this makes you more confused about the issue. It seems, obtaining this additional information diluted your original calculated first stab at the problem. Second. You are drawing from an urn with 10 blue and 90 black lottery tickets. Your initial assessment of the probability of drawing a blue ticket is .1. Then, you learn that the proportion of the tickets at the top is somewhere between .2 and 1. You acquired new evidence, but your evidence became imprecise. In both cases, it seems intuitive that the weight of evidence should decrease, as the evidence becomes less telling.

CITE B. Weatherson. Keynes, uncertainty and interest rates. Cambridge Journal of Economics, 26 (1): 47-62, 2002.]

1.3 Hamer's weight of evidence

1.4 Good's weigh of evidence and the information value

One notion in the vicinity also called *weight of evidence* has been introduced by Good [CITE PROBABILITY AND THE WEIGHING OF EVIDENCE 1950]. Let $W(H:E)$ be the Good's weigh of evidence in favor of H provided by E (if we want to explicitly conditionalize on some background knowledge K, we write $W(H:E|K)$). One assumption about W taken by Good is as follows:

pays attention to different values of different items of evidence, which is better than just counting or supersets

(Function) "It is natural to assume that $W(H:E)$ is some function of $\mathsf{P}(E|H)$ and of $\mathsf{P}(E|\neg H)$, say $f[\mathsf{P}(E|H),\mathsf{P}(E|\neg H)]$. I cannot see how anything can be relevant to the weight of evidence other than the probability of the evidence given guilt and the probability given innocence." [cite Good 1985 p 250]

The other two are:

(Independence) $\mathsf{P}(H|E)$ should depend only on the weight of evidence and on the prior: $\mathsf{P}(H|E) = g[W(H:e),\mathsf{P}(H)]$.

(Additivity) $W(H:E_1 \wedge E_2) = W(H:E_1) + W(H:E_2|E_1)$

The three conditions can be simultaneously satisfied by only one function (up to a constant factor), which leads to Good's definition of weight of evidence:[3]

$$W(H:E) = \log\frac{\mathsf{P}(E|H)}{\mathsf{P}(E|\neg H)}$$

The natural question that arises is the extent to which Good's weight satisfies the desiderata related to Keynes' notion of weight. First, let us think about weight increase with sample size. If in an experiment the observations $E_1,\dots,E_K$ are independent given H and independent given $\neg H$, the resulting joint likelihood is the result of the multiplication of the individual likelihoods, and so the resulting joint weight is the result of adding the individual weights.

For example, suppose a die is selected at random from a hat containing nine fair dice and one loaded die with the chance $1/3$ of obtaining a six. The initial uniform distribution gives you weight of evidence for the die being loaded of $log_{10}(.1)$, that is -1 (Good and Turing would say, it is -10 db). Now, every time you toss it and obtain a six, you gain $log_{10}(\frac{1/3}{1/6}) = log_{10}(2)$, that is 0.30103, and every time you toss it and obtain something else, the weight changes by $log_{10}(\frac{2/3}{5/6}) = log_{10}(.8)$, that is -0.09691. Let us inspect the weights in db (that is, multiplied by 10) for all possible outcomes of up to 20 tosses (Figure 1).

Two facts are notable. (1) Weight can drop with sample size: for instance the weight for 4 others and 5 sixes is 1.2db, and it is .2db for 5 others and 5 sixes. (2) Weight can drop while the sample size increases even if the proportion of sixes remains the same. For instance, if none of the observations are sixes, the weights go from -10 to -19.7 as the sample size goes from 0 to 10. Less trivially, the observation of one six in five leads to weight of -10.9, while the observation of two sixes in ten tosses leads to weight -11.7. That is, (Monotonicity), (Completeness), (Weak increase) and (Strong increase) all fail for Good's measure.

[3] To be fair, logarithms of the ratio of posterior odds to prior odds have been used Jeffrey in 1936, [CITE] and the use of logarithm to ensure additivity has been suggested by Turing [CITE 1950 o 63]. Good's measure differs from Jeffrey's by taking the ratio of likelihoods rather than odds. In fact, the former ratio is identical to $O(H|E)/O(H)$, the ratio of conditional odds of H to the prior odds of H.

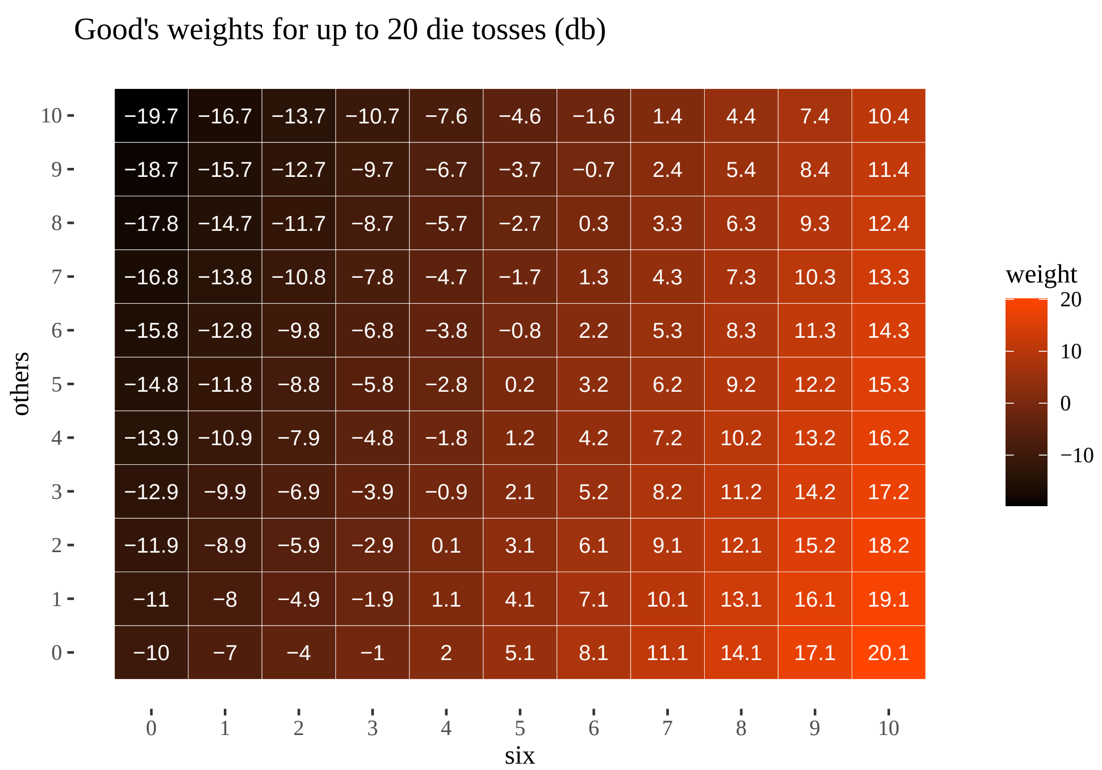

Figure 1: Good's weights in dbs, rounded, for all possible outcomes of up to 20 tosses of a die randomly selected from 10 dice nine of which were fair, and one is 1/3 loaded towards six. H ='the die is loaded'.

Moreover, there is a conceptual difficulty in the neighborhood. Suppose you are trying to ascertain the bias θ of a coin, but you do not restrict yourself to two hypotheses as in the dice example, but rather initially take any bias to be equally likely. For each particular hypothesis $\theta = x$ and any set of observations E you can use the binomial distribution to calculate $\mathsf{P}(E|\theta = x)$. But to deploy Good's definition, you also need $\mathsf{P}(E|\theta \neq x)$, which is less trivial, as now you have to integrate to calculate the expected probability of the evidence given an infinite array of possible values of y. Suppose you have no problem calculating such items. Now imagine you observe 10 heads in 20 tosses. The question 'how weighty is the evidence' makes no sense here, as Good's weight needs a hypothesis (and its negation) to be plugged in. For this reason, in such a situation, we can at best talk about a continuum of Good's weights, one for each particular value of θ.

- compare to pointwise mutual information
- evaluate in light of the desiderata

1.5 Weight and completeness

A question similar to "how weighty is the evidence" is "how complete is it"? These are conceptually different: the former asks about how much information pertinent to a given hypotheses the evidence provides, or, about the amount of evidence relevant to that hypothesis, the latter seems to suggest a comparison to some ideal list of what such items would be needed for the evidence to be complete. While we think that these notions, albeit related, should be clearly distinguished, the distinction has not always been made clearly in the literature, starting with Keynes himself, who suggested that in their evaluation of the evidence an agent should consider "the degree of completeness of the information upon which a probability is based." [TP p. 345] REF

This picture of ideal-list-of-evidence-relative notion of weight has been explored by [CITE FEDUZI 2010]. Let us first present the view following [CITE FEDUZI 343], Ω stands for the set of all items CITE NANCE HERE? of possible evidence relevant for estimating the probability of the hypothesis H. Let K be the agent's knowledge, the set of items of evidence already obtained by the agent, $K \subseteq \Omega$. Then her relevant ignorance is $I = \Omega \setminus K$.

Then, Feduzi, following [CITE RUNDE 1990], proposes to define the weight of information E provides about H, $V(H/E)$ as follows:

$$V(H/E) = \frac{K}{K+I}. \tag{Vdiv}$$

While literally it does not make sense to divide sets by sets, we might charitably interpreting Feduzi as using Ω, K and I the symbols ambiguously, standing for both the sets of items of evidence, and the amount of relevant information that the sets contain. The obvious difficulty is that it is not a successful explication (at least not yet), as we are not told how to get K and $K+I$ as numerical values to be used in the division. But however we get them, let us see whether (Vdiv) can result in any insights.

First, one advantage of the completeness approach is that the resolution of the stopping problem is more or less automatic: the agent should make the decision if the evidence is complete, and should collect more evidence if it is not. Later on, when discussing Nance's approach to the notion, we will see a complication: obtaining further evidence might be practically unfeasible, and so it makes sense to distinguish ideal completeness from reasonable completeness and base the practical stopping rule on the former. For now, we put this complication aside.

Second, it might be the case that obtaining further evidence while providing more information results in the decrease of weight. Here is an example illustrating this due to Feduzi [CITE 345]. Joan in her research tries to establish who is the most quoted author in the literature on decision theory under ambiguity. Ω is the set of all n papers (though of as items of evidence $E_1, \dots, E_n$). K_0 contains the m papers that Joan inspected so far $(E_1, \dots, E_m$, $m < n$. I_0 is the set of papers she did not look at yet, $\Omega = K_0 \cup I_0$. However, Joan is aware only of a part of Ω, the papers in the field she believes exist, S. Thus, her objective ignorance, I_0, and her subjective ignorance, $I_S = S - K$, diverge, as she underestimates the amount of papers that she has not yet encountered. Joan's assessment of weight is going to be $K/_{K+I_s}$. Say Joan formulates a hypothesis, H: "Ellsberg is the most highly cited author in the ambiguity literature" and that she is quite confident that the papers she had not looked at yet would not significantly affect the probability of H. She thinks she has read enough, say $\mathsf{P}(H|K) = .7$ and $V(H/K) = .8$. Then, she looks at another paper, somewhat increasing K, but that paper contains reference to many papers she has not heard of in journals that she has not heard of, thus increasing her estimation of S quite a lot— the ultimate impact of the new evidence is a drop in weight as the denominator in (Vdiv) will grow much more than the numerator. Thus, (Monotonicity) can fail on this approach. check crossref

However, even putting the conceptual difference between weight and completeness that we have already mentioned aside, there are concerns about using degrees of completeness as our explication of the notion of weight of evidence.

To start with, we have not really explicated the notion of the amount of evidence employed in (Vdiv). Sure, we could simply count propositions. One simple strategy, to be used if we do not want to use (Vdiv) would be to simply count the relevant propositions included in the evidence—this would validate (Monotonicity) . Another strategy along these lines would be to assign sizes to sets of propositions and use these as numbers in (Vdiv), invalidating (Monotonicity) in the process. Either way, the strategy is not viable, as it is too syntax-sensitive. Different propositions, intuitively, can contain hugely different amount of nevertheless relevant information , and the individuation of propositions is too arbitrary a matter to take such an approach seriously. On one hand, without some measure of assigning numerical values to sets of evidence, we have no way to deploy (Vdiv). On the other hand, if we could meaningfully assign numbers expressing the "amount of evidence" prior to any application of (Vdiv), there are no clear reasons why we should take these number to express weights of evidence, especially given the second concern with the completeness approach.

check crossref

is this trivial or do we need an example?

So the second difficulty is that on this approach the weight of evidence becomes very sensitive not only to what the actual evidence is, but also to what an ideal evidence in a given case should be. unless as clear and epistemologically principled guidance as to how to formulate such ideal lists is available, this seems to open a gate to arbitrariness. Change of awareness of one's own ignorance, without any major chance to the actual evidence obtained, might lead to overconfidence or under-confidence in one's judgment. Moreover, it is not clear how disagreement about weight arising between agents not due to evidential differences, but rather due to differences in their list of ideal items of evidence should be adjudicated.

1.6 Skyrms and resilience?

-
- relation to law Davidson Pargetter 1986, perhaps Nance, who else?

1.7 Evidential probability and weight

[PEDDEN 2018] follows a suggestion from [KYBURG 1961] He proposed using the degree of imprecision of the intervals in his probability system called Evidential Probability (EP). The key idea in EP is that evidential probabilities should be imprecise, and so accordingly an evidential probability function EP is of the form $\mathsf{EP}(H|E \wedge K) = [x, y]$, where the right-hand is the closed interval expressing the objective degree of support that $E \wedge K$ provide for H.

REF

REF Kyburg. Probability and the Logic of Rational Belief. Wesleyan University Press, Mid- dletown Connecticut, 1961 and H. E. Kyburg and C. M. Teng. Uncertain Inference. Cambridge University Press, Cambridge, 2001.

How is this interval to be determined, though? Kyburg's proposal is the following, if the hypothesis is about a single object o and a predicate P. For the reference classes to which o is known to belong and for which K contains frequency information (possibly imprecise, in the interval form) for objects with P, enumerate the corresponding frequency statements, $r_1, \dots, r_n$. Now, you are facing a reference class problem. Apply sequentially the following rules:

(Sharpening by richness) If r_j conflicts with r_i and r_i has been obtained from a marginal distribution while r_j from a full joint distribution, ignore r_i.

As the formulation might be somewhat cryptic, let us illustrate the recommendation with an example.

make sure these examples are better made sense of the HOP way!

(Richness example) suppose you are drawing a card from one of two decks of cards, $H :=$ 'you will draw the Ace of Spades'. You know that Deck 1 is a regular deck, and Deck 2 is a regular Deck with the Ace of Spades removed. First you toss a fair die and use Deck 1 if the die lands on 1 or 2, and use Deck 2 otherwise. You have at least two frequencies to consider:

- The frequency of Aces of Spades in the total number o cards ($1/103$), which is your marginal-distribution-based-probability.
- The one obtained by using the information about the die, and about the frequencies in the decks. There is probability $1/3$ of using Deck 1 which is the only deck containing the card, in which the probability of drawing it is $1/52$, so the probability to be used is $1/3 1/52 = 1/156$.

One can easily observe that $1/103$ simply is not the probability of drawing the Ace of Spades in the setup. After all, we are not drawing a random card from the joint decks, but have to factor in the uneven probabilities of the decks being chosen, and once we do so, the correct probability is $1/156$. The second strategy is this:

(Sharpening by specificity) If among the remaining intervals r_j conflicts with r_i and r_j is a proper subset of r_i, choose r_j over r_i.

This mirrors the idea that one should use more specific information. The third rule is:

(Sharpening by precision) If there is a single interval that is a proper sub-interval of every other interval, this is the evidential probability. Otherwise, the evidential probability is the shortest possible cover of these intervals.

With this system in the background, [PEDDEN 2018 681] proposes the following definition of the weight of the argument for H given E and K, where $\mathsf{EP}(H|E \wedge K) = [x, y]$:

$$\mathsf{WK(H|E \wedge K)} = 1 - (y - x) \tag{WK}$$

That is, the weight of the evidence is the spread of the evidential probability, transformed to scale between 0 and 1, reaching 1 when the spread is 0 and 0 when the spread is 1.

How is this approach to be applied to examples such as the one by C. S. Peirce (recall: drawing balls with replacement from an urn, with observed frequency of white balls .5, in one scenario the sample size is 1000 in another it is 2)?

[PEDDEN 2018 686] proposes the following analysis, of an example analogous to that by Peirce. You are drawing from an urn full of black and red beans (the proportion is unknown). First, abbreviations:

H 49.5 – 50.5% of the beans are red.
E_1 2 sampled beans are red.
E_2 3000 sample beans are red.

Further, imagine you have enough information to calculate that between 2% and 100% of two-fold samples of any large finite population will be matching samples, that is they will match the population with a margin of error of 1%. Then $\mathsf{EP}(H|E_1 \wedge K) = [.02, 1]$. Similarly, Pedden invites us to suppose that we can calculate the relative frequency of 3000-fold samples that match any large finite population within a margin of error of 1% to somewhere between 72.665% and 100%, so accordingly $\mathsf{EP}(H|E_2 \wedge K) = [.72665, 1]$. Then, the corresponding values of WK are .02 and .72665 (this is because in both cases $y = 1$ and $1 - (1 - x) = x$).

What are we to make of this? Are imprecise probabilities promising when it comes to the explication of weight of evidence? Some progress has been made, but note the following limitations.

- The edges of the intervals are what contributes to WK. They are highly sensitive to the choice of the margin of error, but what margin of error to choose and why remains a mystery, and what margin of error has been chosen does not function anywhere in the EP representation of uncertainty.
- Relatedly, it may easily happen that for two different distributions the 1% intervals will be identical while the 78%$ intervals will not. Such differences will obviously not be captured by the 1% margin of error intervals.
- The calculations of such intervals might be easy for simple combinatorial cases, but it is far from obvious how similar intervals are to be obtained for more complicated real-life cases. Emphatically, classical statistical confidence intervals are not ranges within which the true parameter lies with a certain probability (and if you interpret confidence intervals this way, you behave as if you were running a Bayesian reasoning with a uniform prior, which is often unjustified and prone to over-fitting).

Before we abandon the idea, though, let us know that over the last 30 years we have observed a revival of imprecise probabilities, and so it is only fair that we should take its most recent versions for a ride. Hence our next interest: imprecise probabilism, its motivations, the difficulties it runs into.

1.8 Imprecise probabilities and weight

The point of departure for imprecise probabilism (IP) is the precise probabilism (PP), which we are already familiar with. On the latter view, a rational agent's uncertainty is to be represented as a single probability measure. **Imprecise probabilism**, in contrast with PP, holds that an agent's credal stance towards a hypothesis H is to be represented by means of a set of probability measures, called a representor, $\mathbb{P}$, rather than a single measure P. The idea is that the representor should include all and only those probability measures which are compatible with this evidence. For instance, if an agent knows that the coin is fair, their credal stance would be captured by $\{\mathsf{P}\}$, where P is simply a probability measure which assigns .5 to H. If, on the other hand, the agent knows nothing about the coin's bias, their stance would rather be represented by means of the set of all probabilistic measures, as none of them is excluded by the available evidence. Note that on IP it is not the case that the set represents admissible options and the agent can legitimately pick any precise measure from the set. Rather, the agent's credal stance is essentially imprecise and has to be represented by means of the whole set.[4] The literature contains an array of arguments for IP. Let us take a look at the main ones.

- PP does not seem appropriately evidence-responsive, especially when evidence is limited. Following PP, in Peirce's example, the agent's uncertainty about W :="the next drawn ball is going to be white' is .5 no matter whether you have drawn two balls one of which was white, or a thousand balls, five hundred of which were white.
- Indifference is not sensitive to sweetening (improving the chances of H only slightly), while PP predicts such sensitivity. For instance, if you do not know what the bias of a coin is, learning that it now has been slightly modified to increase the probability of heads by .001 will still leave you unwilling to bet on heads in a bet that would've been fair if the actual chance of H was .5 and not .001.

[4] For the development of IP see (Fraassen, 2006; Gärdenfors & Sahlin, 1982; Joyce, 2005; Kaplan, 1968; Keynes, 1921; Levi, 1974; Sturgeon, 2008; Walley, 1991), (Bradley, 2019) is a good source of literature.

- PP has problems representing complete lack of knowledge. Suppose you start tossing the coin starting with knowing only that the coin bias is in $[0, 1]$ and then observe the outcome of ten tosses, half of which turn out to be heads. This is some evidence for the real bias being around .5. How do you represent your stances before and after the observations? If you deploy the principle of insufficient evidence, you start with $\mathsf{P}_0(H) = .5$ and end with $\mathsf{P}_1(H) = .5$, as if nothing changed. If you do not deploy the principle of insufficient evidence, what do you do?
- PP has problems with formulating a sensible method of probabilistic opinion aggregation Stewart & Quintana (2018). A seemingly intuitive constraint is that if every member agrees that X and Y are probabilistically independent, the aggregated credence should respect this. But this is hard to achieve if we stick to PP (Dietrich & List, 2016). For instance, a *prima facie* obvious method of linear pooling does not respect this. Consider probabilistic measures p and q such that $p(X) = p(Y) = p(X|Y) = 1/3$ and $q(X) = q(Y) = q(X|Y) = 2/3$. On both measures, taken separately, X and Y are independent. Now take the average, $r = p/2 + q/2$. Then $r(X \cap Y) = 5/18 \neq r(X)r(Y) = 1/4$.

One key difference between Kyburg's EP and IP is that on the latter we use sets of probability measure instead of intervals. This makes the approach not only more general (as now, for instance, the resulting probabilities of a proposition in question do not have to form a closed interval), but also provides a more general and less idiosyncratic picture of learning from evidence, that is a a natural extension of the classical Bayesian approach. When faced with new evidence E between time t_0 and t_1, RA's representor should be updated point-wise, running the standard Bayesian updating on each probability measure in the representor:

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

1.9 Troubles with imprecise probabilism

- IP has no means of distinguishing the situation in which you are about to toss a coin whose bias is either .4 or .6., and the one in which you are about to toss a coin whose bias is either .4 or .6, and bias .4 is three times more likely than .6.
- IP give wrong comparison predictions (Rinard, 2013). Suppose you know of two urns, GREEN and $\mathsf{MYSTERY}$. You are certain GREEN contains only green marbles, but have no information about $\mathsf{MYSTERY}$. 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 $\mathsf{MYSTERY}$ will be green (M). In line with how lack of information is to be represented on IP, for each $r \in [0, 1]$ your representor contains a P with $\mathsf{P}(M) = r$. But then, it also contains one with $\mathsf{P}(M) = 1$. This means that it is not the case that for any probability measure P in your representor, $\mathsf{P}(G) > \mathsf{P}(M)$, that is, it is not the case that RA is more confident of G than of M. This is highly counter-intuitive.
- While IP, does not seem to have a problem reprentic complete lack of information, the way it is represented makes agents susceptible to belief inertia (Levi, 1980). Recall the example in which you start tossing a coin starting with knowing only that the coin bias is in $[0, 1]$ and then observe the outcome of ten tosses, half of which turn out to be heads. On IP your initial credal state is to be modeled by the set of all possible probability measures over your algebra of propositions . Once you observe the ten results, each particular measure from your initial representor gets updated to a different one that assigns a value closer to .5 to Heads, but also each measure in your original representor can be obtained by updating some other measure in your original representor by updating on the evidence. Thus, if you are to update your representor point-wise, you will end up with the same representor set.

1.10 A second-order approach to uncertainty

1.11 Information-theoretic weight of evidence

1.12 Completeness tends to improve weight

1.13 Weight tends to improve accuracy

Here is a question asked by [COHEN 1986 TWELVE p. 276]: is it worth while knowing the weight of an argument without knowing its probability? In our terminology, questions inspired by Cohen's are: what's the point of weight considerations if we already have the distributions? Can weights be put to use if we do not have the distributions?

2 Literature to discuss

Kasser, 2016, Two Conceptions of Weight of Evidence in Peirce's Illustrations of the Logic of Science [COVERED]

Feduzi, 2010, On Keynes's conception of the weight of evidence COVERED

Cohen 1986, Twelve Questions about Keynes's Concept of Weight [COVERED]

Pedden, William 2018, Imprecise probability and the measurement of Keynes' weight of arguments [GET BACK TO INERTIA, DILUTION ETC.]

Levi 2011, the weight of argument [DOWNLOADED]

Skyrms 1977 resiliency, propensities [DOWNLOADED]

Skyrms causal necessity, chapter on resilience [DOWNLOAD]

Synthese 186 (2) 2012, volume on Keynesian weight [CHECKED, NOT MUCH ON WEIGHT ACTUALLY, NO NEED TO READ]

Good, weight of evidence, survey

Good, PROBABILITY AND THE WEIGHING OF EVIDENCE

David Hamer, Probability, anti-resilience, and the weight of expectation [READ]

William Peden, Imprecise Probability and the Measurement of Keynes's "Weight of Arguments"

Runde, Keynesian Uncertainty and the weight of arguments [DOWNLOADED]

Weatherson, 2002, Keynes, uncertainty and interest rates [DOWNLOADED]

Jeffrey M. Keisler, Value of information analysis: the state of application

Edward C. F. Wilson, A Practical Guide to Value of Information Analysis

Joyce JM (2005) How probabilities reflect evidence.

Kyburg. Probability and the Logic of Rational Belief. Wesleyan University Press, Middletown Connecticut, 1961

H. E. Kyburg and C. M. Teng. Uncertain Inference. Cambridge University Press, Cam- bridge, 2001.

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.

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

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

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

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

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

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

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

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.

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

Stewart, R. T., & Quintana, I. O. (2018). Learning and pooling, pooling and learning. *Erkenntnis*, *83*(3), 1–21. https://doi.org/10.1007/s10670-017-9894-2

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

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