Comments for Referees

Reviewer #1

I am happy that the author has responded to my previous comments listed in their response. A couple of mostly niggly points on the revised manuscript:

(1) p. 2: Isn't it more an issue of rationality rather than numerical effectiveness? Perhaps delete: "numerically most effectively represented". I agree that the term "numerical effectiveness" is misleading and ambiguous. I wanted to keep the door open to qualitative representations of partial beliefs that run parallel to a representation in numbers. I have changed the passage to "the best quantitative model of partial beliefs is such that partial beliefs correspond to probabilities"

(2) p. 5: Perhaps "agree" is better than "aligned" at the bottom. Also, perhaps "The information theorist is the better Bayesian" or "information theory is the better Bayesianism" is better shortly after? I have incorporated both of these suggestions in the revision.

(3) p. 15: "The conclusion follows." Is it worth restating the conclusion here? I have deleted this sentence in the revision.

(4) p. 26: "Unfortunately for Pettigrew..." This was a little confusing. Perhaps rephrase: "Unfortunately for Pettigrew, the Log score also has a claim to uniqueness, namely, …" I have changed the passage to "Unfortunately for Pettigrew, the Log score also has a claim to uniqueness. It is the only proper scoring rule which fulfills locality."

(5) p. 41: "They are roughly aligned..." Perhaps inserting "respectively" somewhere here would help the reader. Yes! I have inserted the missing "respectively."

(6) p. 62: "degree of confirmation theory". Would simply "confirmation theory" be better? I deleted "degree of" for the revision.

(7) p. 69: Perhaps replace "classification" for "clarification"? I have deleted this expression altogether and replaced the parenthetical expression by "It would help, for example, to know that all reasonable non-commutative difference measures used for updating are ill-behaved."

Reviewer #2

Summary

This revised manuscript is improved and may possibly be accepted after carrying out major revisions. Given the number of comments and the severity of some of them, I can see an argument for rejecting this manuscript; see below for details.

Major Comments

The major problem I have with this problem is that there are a large number of things in it which are not quite right – to my eyes.

I have made substantial changes to the manuscript following the recommendations of the referee. The major changes are with respect to the following items:

- Clarify the relationship between affine and convex constraints
- Change the setup for the probability space, following Predd et al.
- The Log score fulfills LOCALITY only on probability functions
- Fill out and clarify the proof for symmetry
- Change the treatment of continuity, as LP-“conditioning” isn’t even continuous with standard conditioning in standard conditioning cases

Minor Comments

Page 2: “the updating scenario permits it”? I have replaced this confusing expression by “when possible”

“if there is no further information, mutually disjoint and jointly exhaustive events are equiprobable” – that’s too simplistic. Surely, for $\Omega = \{\omega_1, \omega_2, \omega_3\}$ the PoI does not recommend that $P(\omega_1) = P(\omega_2 \cup \omega_3)$. Quite right! I have replaced the passage by the following: “if there is no further information, then mutually disjoint and jointly exhaustive events distinguishable by name only are equiprobable.”

“report” ? I was trying not to commit to prediction or other types of belief representation, but I agree that as it stands the expression is confusing. I have tried to improve the situation as follows: “Let the agent’s ‘report’ over these three outcomes be $c=(c_{1},c_{2},c_{3})^{\intercal}$. The report may represent a partial belief or an attempt at prediction.”

Some of the c i may be zero, right? It's worth stating this explicitly. I was hoping that the expression "non-negative real numbers" was sufficient to allow for zeroes.

"she has no other concerns" – why not say that L captures all the agent cares for? I have changed this passage to "the agent does not care for anything outside of what the loss function is able to capture."

Please, do use the word strict/strictly everywhere. It is too confusing for readers otherwise. I followed this recommendation and replaced "proper/propriety" by "strictly proper/strict propriety" whenever applicable.

Page 4, can we not visualise non-Euclidean space? I have changed the clause to "which is more readily accessible to
intuition."

Page 5, there seems to be a word missing in the last sentence (after the first appearance of 'Bayesian'). This is a sentence that needed to be improved in response to the first reviewer, so the problem has been fixed.

Page 6, cross entropy updates can cope with more complicated constraints, not just affine constraints. Convexity of the feasible region is normally required to guarantee uniqueness of the solution. Thank you for pointing this out; I was only dimly aware of the relationship between affine and convex constraints, nicely illustrated in Paris's *Uncertain Reasoner's Companion* in Proposition 6.1 on page 66. I have changed the passage to the following: "The more general updating situations include affine constraints (as in the notorious Judy Benjamin example), of which Jeffrey-type updating scenarios are a special case, and even more complicated constraints where the evidence only allows for a convex region of credence functions."

"The Log score is asymmetric, but unlike the Brier score not unique among its asymmetric peers" – these (non-)uniqueness claims require more hedging. I have made the statement more precise by reformulating "The Log score is asymmetric, but there are other asymmetric strictly proper scoring rules."

List A and List B: It is preferable to only have the explanation of the feature in the list. All evaluations and examples should appear outside the list. I have made the corresponding adjustments and moved evaluations and examples to the main text, except in List B where I provide one example per list item which makes the abstract property more immediately intuitive (and there is some symmetry and organization to the list in so far as each list item gets an example).

List A, no substantiation is made for the "other scores" column. Given that his column plays no role in this manuscript, I suggest to delete it. I am happy to do this, but want to point out that (a) the 'other scores' column clarifies that there are no other scores beside the Log score and its close relatives which will fulfill all plausible desiderata; (b) the 'other scores' column illustrates that while the Brier score is uniquely symmetrical, the Log score is not uniquely asymmetrical, and while the Log score uniquely fulfills locality, the Brier score does not uniquely violate locality.

Page 8, the set up of the probability space is unsatisfactory. Why not simply say that “the finite set of elementary events Ω is generated by n binary variables”? The original (pre-revision) event space \script{A} did not only contain events such as A \cup B and complement{A} \cup B, but also their unions. This induced logical relations between the events in \script{A} which then posed constraints on credence functions that qualify as probabilities. The space of credence functions which violated the logical relations (and which may or may not have fulfilled Kolmogorov’s axioms of probability) was vast -- \math{R}^{256} for three binary variables. I thought it was important to include this information to avoid the objection “But what about credence functions which violate the logical relations (and possibly also probabilism)? Might they outperform probabilistic credence functions that obey the logical relations?” This was a real concern, and I included the complicated set up in order to address it. The referee disagrees that my setup did any of this work, and I agree. There is no proof of the claim in the pre-revision paper that “all arguments defending probabilism in this paper do not only justify probabilistic credence functions over non-probabilistic ones that obey the logical entailment relationships, but ad fortiorem also over non-probabilistic ones that disobey the logical entailment relationships”—the referee even believes that this claim is false. As a result, I have to settle for now with a more modest claim for the paper and restrict myself to credence functions that are logically coherent and determined by the numbers they assign to mutually exclusive and collectively exhaustive events. I used Predd/Seiringer/Lieb/Osherson/Poor/Kulkarni’s setup in the revision, who also restrict their attention to this type of credence function.

The negation symbol “¬” is normally not a superscript. When logical notation (¬) is used, then one normally uses ∧ instead of ∩. I agree that this is inconsistent. I have consistently changed the notation to set notation, as I am taking events to be sets.

Page 9/10: This just means that you accept the additivity axiom. At this point [at the very latest], you need to give a precise definition of the credence functions you are working with:
$C := \{c : P\Omega \to [0, \infty) \mid \text{SUM\_ } \omega\in\Omega\ c(\omega) > 0 \ \&\ F \cap E = \emptyset \Rightarrow c(F)+c(E) = c(F \cup E)\}$?? I have replaced my more idiosyncratic setup by Predd/Seiringer/Lieb/Osherson/Poor/Kulkarni’s setup in the revision, so this concern has been addressed. See referee’s comment on page 8 and my response above.

“All arguments defending probabilism in this paper do not only justify probabilistic credence functions over non-probabilistic ones that obey the logical entailment relationships, but ad fortiorem also over non-probabilistic ones that disobey the logical entailment relationships.” No. See my later comments on logarithmic scores. I have replaced my more idiosyncratic setup by Predd/Seiringer/Lieb/Osherson/Poor/Kulkarni’s setup in the revision, so this concern has been addressed. See referee’s comment on page 8 and my response above.

The first paragraph of Section 2.2 is a mess. Since the set of probability functions is convex, its convex hull is equal to this set. This set is a proper subset of the set of credence functions. For a credence function c which is also a probability function, one can of course chose $p = c$. This result holds much more widely for continuous strictly proper scoring rules and credence functions which do not satisfy additivity, see [2]. [3] only considers probabilistic credences. I have replaced the passage in the revision by the following more accurate and detailed version: "For any vector c in the vector space of credence functions, there is a vector p in the set of probability functions which is closer to each possible world than c, where closeness is evaluated in terms of a suitable measure of closeness, for example a continuous strictly proper scoring rule (Predd et al., 2009, call continuous strictly proper scoring rules 'proper scoring rules' and use the corresponding Definition 2 to prove de Finetti's result in Theorem 1 on page 4788). If c is not a probability function, then the vector p is strictly closer to each possible world than c. If c is a probability function, then one trivially chooses $p = c$."

delete bracket on evidence. Recommendation implemented.

Classical information theory only considers P probabilistic credences. Their logarithmic loss function is an expectation, $P (\omega) \log(Q(\omega))$. According to this expected loss, $Q(\omega) = 1$ for all $\omega \in \Omega$ is the best possible choice. Since standard information rules out such credences by stipulation, this is no internal problem for them. The relation between their logarithmic scoring rule and the logarithmic scoring rule popular in formal epistemology circles [which is strictly proper also for non-probabilistic credences] is discussed in [1]. In your equation 12, this logarithmic scoring rule is not local for non-probabilistic credences. This is absolutely correct and an astute observation. I have added a reference to Landes and the following explanation to the paragraph: "Note that the Log score violates \textsc{locality} on $\mathcal{D}\setminus\mathcal{P}$, so that arguments using the unique characteristic of the Log score to fulfill \textsc{locality} presupposes an independent argument for probabilism (see Landes, 2015)."

Add 'p.' to the Hendrickson bracket. I also noticed how awkward "1918" is as a page number, but for consistency's sake left out the 'p.' — I am easily persuaded and have added the 'p.' for clarification.

Page 14: add that close relatives have the same extrema. I have added the following remark: "They do not differ from them in terms of optimization (i.e. their extrema are equivalent)."

Lemma 3.2: You should define $\leq$ for functions. Presumably, it's some point-wise dominance notion. I have added a definition of $f_1<f_2$ in terms of point-wise dominance.

Lemma 3.4: Do you mean $(\nabla H)-1 = \nabla(H*)$ or $(\nabla H)-1 = (\nabla H)*$? I mean $\nabla(H*)$ and have disambiguated accordingly in the revision.

"Briar" – Brier (twice) I fixed the typo. Thanks for catching it.

"Lemmas 3.4, 3.5, and 3.6 establish theorem 3.1" – how, why? I have decided to replace this proof with a better one based on Reinhard Selten's work. My proof differs in several interesting respects from Selten's, but the basic idea is his.

Page 22: A scoring rule cannot embrace an epistemic norm. I have improved this passage to read “I describe a scenario where one way for an advocate of the Brier score to escape a dilemma described there is to embrace regularity.”

The paragraph on confirmation measures is under-developed and best omitted. As recommended by the referee, I have deleted the paragraph on confirmation.

Last paragraph of Section 4.1: Agreed; but so what? You are really comparing apples to oranges here. I have added an explanatory note: “The greater penalty for Tatum may strike one as counterintuitive, but it is a natural consequence of a scoring rule violating LOCALITY.” I don’t understand the apples to oranges comparison. Tatum’s forecast is 12% for colour 1; Casey’s is 10%. Colour 1 turns out to be true. It is initially counterintuitive that Tatum is penalized more severely than Casey by the Brier Score (the Log Score, obeying LOCALITY, penalizes Casey more severely than Tatum).

Page 25 on Joyce: No, Joyce shows no such thing. The result you refer to is proved in [2] for continuous and strictly proper scoring rules. The result of Joyce is properly stated in 5.1. I have changed the wording of this passage to make my claim more accurate and to give proper credit to Predd et al.

Page 27: I found the last part of the second to last paragraph opaque. I agree. I was hoping to illustrate the argument in the next paragraph. Sometimes there is an independent reason for uniqueness (a pistol pointed at you to identify a unique candidate)—my argument is that the need for uniqueness cannot be added to the list of reasons supporting the uniqueness of the candidate. If Peter, Paul, and Mary are my children, and I tell you (with a pistol on my chest) that Peter is my favourite child, then the list of reasons why he is my favourite child includes his cleverness, his haircut, and his disdain for Pokemon cards, but it does not include the fact that I chose him when coerced to identify a favourite child.

The next paragraph should be connected to the rest of the text. See me response to the last comment.

Theorem 4.4: Again, this only holds for probabilistic credences. How to save some kind of locality notion for logarithmic scoring rules which are strictly proper for all credence functions is discussed in [1]. Excellent point and not something I had considered. I inserted the following explanatory paragraph (thanks to this referee), “Note, however, that the Log score only fulfills LOCALITY if the domain on which LOCALITY is evaluated is probabilistic credences. Jürgen Landes discusses how to save some notion of locality for logarithmic scoring rules which are strictly proper for all credence functions (see Landes, 2015).”

How can a scoring rule vindicate conditioning? I have changed the passage as follows: “Given Leitgeb and Pettigrew's assumptions, standard conditioning is vindicated, but Jeffrey conditioning is ruled out.”

Page 32: Sure, but we can also embed into a different space (or even manifold) with a different [e.g., Riemannian or symplectic] metric. I have added the qualification “this geometric relationship is not necessarily Euclidean”—anticipating the more detailed explanation in the next paragraph.

INVARIANCE: that's very vague. I have specified the requirement as follows: "An updating method ought to be partition invariant in the sense that introducing irrelevant subpartitioning does not change the outcome of the update."

[4] does not justify probabilism. That's right. I removed the reference to probabilism. (Does Theorem 7.2 in Landes, 2015, "justify" probabilism in terms of information theory?—not quite, but it creates a strong association between them.)

Page 40: these are different (non-equivalent) metrics which – of course – give different distances and hence make different recommendations. The point (perhaps not so remarkable) is that not only do they give different distances but also change the ordering of who is closer to whom.

(49): GExp A not defined. I was hoping that the detailed reference to Leitgeb and Pettigrew would make the definition unnecessary. It is not used anywhere else in the paper.

5.4.1: LP updating differs from conditionalisation. No continuity argument is required to make this point. The continuity argument is indeed way simpler than in my pre-revision version. My original "parallel sequence of updating scenarios" is unnecessary, as the referee correctly points out. In the post-revision version, however, I have retained the idea of a continuity requirement, this time to express the desire that updating in Jeffrey type updating scenarios should be continuous with standard conditioning in the sense that where standard conditioning applies, LP updating doesn't give us a different result than standard conditioning. As the referee points out, LP updating is not "continuous with standard conditioning" in this sense. This is why the referee calls it "LP updating" rather than "LP conditioning," which is my terminology in the paper. I am open to changing my terminology, although I have not done so in the current version. I have changed and simplified the whole subsection 5.4.1 to reflect the referee's concern.

(58): that's not a probability function. Quite right (a calculation error), but based on the changes to this whole subsection (see previous comment), the algebra has disappeared and this is no longer a problem.

Page 48/49: This only applies within the standard Bayesian framework. LP are silent about how to use their probabilities for decision making. Also, in their approach a finite amount of information can change posterior probabilities from or to zero. Good point. It is ironic that I missed this considering that I have a subsection dedicated to EXPANSIBILITY, which restates the concern of the referee. For example, the probabilities (0, 1/3, 2/3) are LP-updated to (5/24, 13/24, 1/4) for P'(E_3)=1/4. A wide swath of evidence will result in a strictly positive posterior probability for events with a prior zero probability in LP updating. I have deleted the passage making the decision-theory/information-theory claim.

5.4.4: this needs to be better motivated. It does look to me like these agents use different partitions and hence it is not immediately clear why Invariance should apply here. I have added the following to motivate INVARIANCE: "In the following example, Sherlock Holmes and Jane Marple agree on all relevant facts and on their prior probabilities, but LP conditioning leads to a divergence in posterior probabilities."

Page 52: Why would one use standard conditioning in a previous step when one later uses LP updating? In section 6.2 of their paper “An Objective Justification of Bayesianism II” (2010), Leitgeb and Pettigrew vigorously defend standard conditioning. The referee is correct here that LP conditioning is not “continuous” with standard conditioning in a stronger sense than I describe in the paper—I have addressed this problem in my response to the recommendation for pages 48/49. The impression that LP’s paper gives, however, is such that standard conditioning should be applied when possible (when P(E)=1) and LP conditioning otherwise. The referee’s observation highlights the problem with LP conditioning.

Surely, the proper way is to specify the total available evidence, which either pronounces on X_0 or not. Either way, it’s clear how the update should be carried out. There are two steps here. We end up with an event (X_0) whose probability in the first step is diminished to zero. Then additional evidence comes in calling for further updates. These two steps are not meant to be available as ‘total evidence’ in one single step.

Page 55: The first paragraph conflates all sorts of intuitions and metrics. I take this to be referring to the following section: “Near the boundary of $\mathbb{S}^{n-1}$, information theory reflects the horizon effect just as our expectation requires. The problem is near the centre, where some equidistant points are more divergent the closer they are to the middle.” I have deleted the word “equidistant” in the revision, but the problem remains that – as the referee correctly identifies – the whole idea of collinear horizon is based on a conflation of metrics. I am aware of this; in fact, I am currently working on a paper that presents a coordinate-free approach to scoring rules which is less dependent on geometric intuitions. To address how problematic collinear horizon is in this paper (in its defence, it does have the interesting feature of showing that both the geometry of reason and information theory fail the horizon requirement), I have expanded the following passage in subsection 5.4.6: “The way I have formalized the HORIZON and COLLINEAR HORIZON requirement is artificial in the face of the more comprehensive epistemic intuition. COLLINEAR HORIZON conflates divergences and metrics as it is dependent on the Euclidean idea of collinearity and equidistance. In a more integrated account it would be desirable to have these requirements reformulated in a more general fashion; convexity may play a major role in such a reformulation.”

Figure 5: How can it be unclear, whether a point is a midpoint or not? A calculation should make things clear, no? I thought so, too! There is no algebra that I am aware of that would show identity between midpoint and point E. The Lambert-W function must be calculated numerically; using matlab I was not able to find a difference between midpoint and point E, but theoretically speaking they could be so close to each other (while not being identical) that my numerical methods are insufficient to demonstrate the difference.

(70): What if $\beta > \alpha$? Good point. I have added the necessary qualification: “can be expressed using three variables suitably constrained to yield probabilities (for example, $\alpha-\beta>0$).”

C is not right: there are some min{..., 0} missing. True—however, I have excluded those cases by noting earlier “assuming that LP conditioning does not ‘fall off the edge’ as in case (b) in Leitgeb and Pettigrew, 2010ii, page 253.”

Please, do show the algebra. I wrote out the identity in the text of the paper (see revision) and checked again that it nicely reduces to a trivial identity, which it (remarkably) does.

Page 59: Do not use the “simplification”. There’s no point in introducing a symbol which is only used a handful of times. I concur. I implemented the recommendation.

(78): The reasonable thing to require might be to point out that ||b − a|| is symmetric in a and b, while information geometry is not. Hence, collinear horizon should then only require that D KL (B, A) > D KL (B, C). A and B are metrically closer to the centre than B and C; and should therefore be SR-closer to each other than B and C, given that both pairs (A,B) and (B,C) are metrically equidistant. The referee is correct that there was an incongruity in the original text, although I can’t quite align it with the referee’s comment. I have fixed the incongruity as I understand it (for example, “a,b,c” had to be replaced by “p,p’,q’,q’”).

I can’t make much sense out of the caption of Figure 6. The caption is explained at the end of section 5.5.2, but I agree that the whole comparison is a bit far-fetched. I was noticing the similarity in pattern, but I have no explanation for it. I am happy to cut this passage from the paper entirely.

References

[1] Jürgen Landes. Probabilism, Entropies and Strictly Proper Scoring Rules. International Journal of Approximate Reasoning, 63:1–21, 2015.

[2] J.B. Predd, R. Seiringer, E.H. Lieb, D.N. Osherson, H.V. Poor, and S.R. Kulkarni. Probabilistic Coherence and Proper Scoring Rules. IEEE Transactions on Information Theory, 55(10):4786–4792, 2009.

[3] Leonard Jimmie Savage. Elicitation of personal probabilities and expectations. Journal of the American Statistical Association, 66(336):783–801, 1971.

[4] John E. Shore and Rodney W. Johnson. Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-Entropy. IEEE Transactions on Information Theory, 26(1):26–37, Jan 1980.