Entropy **2014**, *16*, 6186-6194; doi:10.3390/e16116186

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Article

What You See Is What You Get

Jeff B. Paris

School of Mathematics, The University of Manchester, Manchester M13 9PL, UK;
E-Mail: jeff.paris@manchester.ac.uk; Tel.: +44-161-275-5880; Fax: +44-161-275-5819

External Editor: Juergen Landes

Received: 23 June 2014; in revised form: 30 October 2014 / Accepted: 4 November 2014 / Published: 21 November 2014

Abstract: This paper corrects three widely held misunderstandings about Maxent when used in common sense reasoning: That it is language dependent; That it produces objective facts; That it subsumes, and so is at least as untenable as, the paradox-ridden Principle of Insufficient Reason.

Keywords: language dependence; insufficient reason; Maxent; uncertain reasoning

1. Introduction

Around the mid 1980s I became fascinated by the idea of Expert Systems (as they were called then). Basically the idea my collaborators and I had was that we would extract lots of knowledge from medical experts and then have the computer use it to do their job, thus making most of the medical profession redundant. Disheartenment soon followed and instead we looked at the conception that this approach was based on, in particular that we humans, or maybe some other sort of agents, might be the possessors of some basic "knowledge", or "qualified belief" set, which could then be completed using some sort of of "Inference Process".

The context we settled on was as follows. We have a finite Propositional Language L, say L has propositional variables $p_1, p_2, \ldots, p_q$ with SL denoting the set of sentences of L, and our agent has some knowledge/belief set K about the probabilities of certain sentences θ_i from SL.

More specifically we assumed K to be in the form of a (satisfiable) set of linear constraints

$$K = \begin{cases} \sum_{i=1}^{m} a_{1i} w(\theta_i) = b_1, \\ \sum_{i=1}^{m} a_{2i} w(\theta_i) = b_2, \\ \qquad \vdots \\ \sum_{i=1}^{m} a_{ki} w(\theta_i) = b_k, \end{cases}$$

on a probability function (w) which the agent's personal probability function is supposed to satisfy (whilst this may seem to offer rather limited possibilities for the sort of knowledge which can be expressed the equality relation in these constraints can be replaced by $\leq$ without changing any of the results we shall need herein). Strictly speaking w here is a variable standing for a probability function on SL, that is a function $w : SL \to [0, 1]$ satisfying, for $\theta, \phi \in SL$:

(P1) $\models \theta \Rightarrow w(\theta) = 1$,

(P2) $\theta \models \neg\phi \Rightarrow w(\theta \vee \phi) = w(\theta) + w(\phi)$.

On the face of it, this might seem like a very special and restricted context. Nevertheless, it is one which is met rather frequently in discussions in Formal Epistemology.

Given such (discrete as opposed to continuous) knowledge bases the question we were interested in was "How should our agent select a personal probability function to satisfy these constraints?" The intention here was that the agent would make this choice in a way which was somehow rational or common sensical *under the assumption that K constituted the sum total of his/her knowledge*. To this end Alena Vencovská and I started collecting/formulating common sense requirements that the *process* of assignment

$$I : K \mapsto I(K) = \text{ probability function } w \text{ satisfying } K,$$

should satisfy, dubbing such an I an "Inference Process".

To our great surprise it turned out that the common sense requirements/principles that we formulated left no choice as to what I could be, the principles forced I to be *Maxent*, *i.e.*,

$$I(K) = \text{ the solution to } K \text{ which maximised the entropy,}$$

see, for example, [1,2]. (Actually we were subsequently accused of fabricating these principles in order to isolate this Maxent solution, but that really was not the case.) Results characterizing Maxent in terms of "Rational Principles" had been obtained before, e.g., by Shore and Johnson, see [3], but these had all assumed the choice of probability function to be the maximum point of some function F, the problem then being to characterize F. We made no such assumption, only that the agent was making *some* choice (Independently, Csiszár in [4] also gives a characterization of Maxent (and Least Squares *etc.*) without this assumption.) In a later paper, [5], the results of [1] were extended to much more general knowledge bases.

Over my time of working on this area I came to the opinion that Maxent is synonymous with "common sense", at least in this limited situation. It rather irked me then when not everyone agreed with me (!) and as a result in 1997 Alena Vencovská and I published [6] to address these (and other) common misunderstandings. Unfortunately, these delusions were not completely exorcized, hence this current attempt to hopefully save some souls from eternal darkness.

2. Maxent and Language Dependence

In [7], and elsewhere, Teddy Seidenfeld (amongst many others. I have chosen this paper because of its especially wide influence) seemingly intends to criticize the Maxent inference process, ME say, for being "Language Dependent", meaning that for the "same" knowledge K expressed in another way, say K', we may have $ME(K) \neq ME(K')$. In other words the probability function satisfying K with the maximum entropy is not the same as the probability function satisfying K' with maximum entropy, despite K and K' "representing the same knowledge" (subsequently there have been various attempts to solve this general "problem", for example [8,9]).

He gives the following example of this. Consider a conventional die with faces marked 1 to 6. Let p_i stand for

$$\textit{The die lands with face } i \textit{ uppermost} \tag{1}$$

He does not spell out what he takes the knowledge to be in this case but apparently it is

$$K = \begin{cases} w\left(\bigvee_{i=1}^{6} p_i\right) = 1, \\ w(p_i \wedge p_j) = 0, \quad 1 \leq i < j \leq 6. \end{cases}$$

In this case we obtain that $ME(K)(p_i) = 1/6$ for $i = 1, 2, \dots, 6$, just what one would surely consider "common sense" on grounds of symmetry. In particular then the "common sense" probability of a 1 comes out to be 1/6.

Seidenfeld now suggests we consider the problem of what probability to give a 1 being uppermost on the next throw with a different partition of cases, namely let $q_1, q_2, \dots, q_{14}$ denote the 14 possible views we could have when the die lands when a "view" is determined by

(1) The score on the uppermost face.

(2) Which of $i < j + k, i = j + k, i > j + k$ holds, where i is the uppermost score and j, k are the scores on the two visible side faces (or just drop k if only one side's face, showing j, is visible).

(There are less than the apparent 6×3 cases because some combinations are not possible.) According to his figures he now takes the knowledge K' to be

$$K' = \begin{cases} w\left(\bigvee_{i=1}^{14} q_i\right) = 1, \\ w(q_i \wedge q_j) = 0, \quad 1 \leq i < j \leq 14. \end{cases}$$

In this case we obtain that $ME(K')(p_i) = 1/14$ for $i = 1, 2, \dots, 14$, just what one would surely consider "common sense" on grounds of symmetry. On the basis of K' alone there is no rational reason for giving different probabilities to any of the p_i and so, since they are disjoint and exhaustive according to K' they should all get probability $1/14$. However, since for only one of these 14 alternatives is it the case that a 1 is uppermost (since then certainly $1 < j$) the probability according to ME on the basis of K' of the die landing with a 1 uppermost is 1/14, so not 1/6.

What is wrong here is that K' does not express "all the knowledge". Once the answer 1/14 comes out extra knowledge, not included in K', about this particular situation is being introduced in order to discredit that answer. The fact that K' derives from a finer partition of the outcomes of throwing the die does not mean that it incorporates more knowledge. Indeed in K' the essential symmetry of the die has been lost.

To make this plain put these two previous examples out of your mind for the moment and consider being told that a 14 sided (though some of these faces may be degenerate, *i.e.*, simply corner points or edges) convex polyhedron, with the faces marked $1, 2, \ldots, 14$, was to be thrown. Let q_i, $i = 1, 2, \ldots, 14$ now stand for it landing with face i bottom-most. In this case the knowledge could, much more reasonably in fact, be represented again by K', and in this case $ME(K')(q_1) = 1/14$ is surely the "common sense" answer. Of course we could now "re-express" the knowledge (so exactly reversing what happened with the 6 sided die) by letting:

p_1 stand for "face 1 bottom-most",
p_2 stand for "face 2 or 3 bottom-most",
p_3 stand for "face 4 or 5 or 6 bottom-most",
p_4 stand for "face 7 or 8 bottom-most",
p_5 stand for "face 9 or 10 or 11 bottom-most",
p_6 stand for "face 12 or 13 or 14 bottom-most".

Now, apparently in accord with Seidenfeld, our knowledge becomes K and we obtain $ME(K)(p_1) = 1/6$, so not $1/14$.

The point of this example is that here we have the same K, K' but a different interpretation and it is now the K' which, according to this interpretation, gives the answer I think most of us would feel was reasonable. The K and K' of course are ignorant of the user's intended interpretation, except in as far as they express some of the constraints which it fulfills.

In short it is a case of "what you see is what you get". K, K' are just uninterpreted sets of constraints and there is nothing there on the page about a die and summing faces *etc.*

Apart possibly from the third of the above examples in each case the constraints which are supposed to sum up the knowledge are lacking. So what happens is that in each of these cases we base a solution on inadequate knowledge and then reveal extra knowledge to ridicule the answer Maxent (in this case) gives.

One might argue here that nevertheless this "extra knowledge" is knowledge and so should be included. If one thought that however, that such knowledge was not irrelevant and so should be included, then one should also include an exactly parallel extension of the original knowledge based not on the numbers 1,2,3,4,5,6 in that order but also on the numbers $\sigma(1), \sigma(2), \sigma(3), \sigma(4), \sigma(5), \sigma(6)$ in that order for each permutation σ of $1, 2, 3, 4, 5, 6$. It is easy to see that that returns us to the original solutions of 1/6. Notice that a similar "fairness" in the last example above also gets us back to the original 1/14 (this point is discussed in detail in [6]).

In short, the examples fail the essential requirement of having EVERYTHING that is known (to the inferring agent) included in the knowledge base we are applying Maxent to. In fact the only one of these that gets near to that requirement of total knowledge is the third example of the 14 sided polyhedron. (A missing piece of knowledge is that the polyhedron must have at least two non-degenerate sides, though in fact adding that would not change the Maxent solution.)

Of course, as these examples show, the same problem would arise with any inference process which respected symmetries, there is nothing special about Maxent here. And if an inference process does not respect symmetries what is left?

The misunderstanding which Seidenfeld and many many others have propagated here and which is held very widely in Philosophy (in my experience) can be summed up as follows: A formal, abstract (*i.e.*, involving an uninterpreted language) method is proposed for drawing a conclusion C from a knowledge base K (in a wide sense of the term, not necessarily as specific as above). An opponent of the method then applies an interpretation to the symbols in K and C, giving say K^I, C^I, and points out, to the discredit of the method, that C^I is a quite ridiculous conclusion to draw from K^I.

An examination of what specific knowledge is being invoked to render this conclusion ridiculous now (almost invariably) shows that said invoked knowledge had been omitted from the original "everything that is known" knowledge base we started with.

But, one may argue, *how can one ever hope, in general, to put down "all that one knows"?* Well, yes, that's certainly a serious objection for the everyday real world use of Maxent, though not an inevitable shortcoming as our third example above perhaps shows. What can be learnt from this latter is that if the "relevant knowledge" can be formulated in the required form then ME is an arguably justified way to infer common sense subjective probabilities from it. Or, in a more advisory, rather than prescriptive, mode, if these inferred probabilities seem contra-common sense then the relevant knowledge (or ME!) needs to be re-evaluated.

Having raised this possibility it opens up the question of how we can ever hope to know what the relevant knowledge is? (This issue is considered in some detail in [6].) Obviously, we did not need to put in any information about, say, my dog, in the knowledge base in the third example so we do seem to have some intuitive idea of what information is relevant but giving a precise prescription seems problematic. For example on what grounds apart from my "gut feeling" do I consider information about my dog irrelevant to the throw of the 14 sided polyhedron?!

Fortunately when we write down a set of constraints, and necessarily have to omit much additional knowledge on the grounds of its perceived irrelevance, others tend share our view on what is irrelevant. For example in the case of a die most of us I imagine would feel that K summed up the relevant knowledge and would be happy to give a (subjective) probability of 1/6 of a 1 landing uppermost. But if you were in, I suppose a rather small, minority who felt that K' summed up all the relevant knowledge then you would opt instead for a probability of 1/14. That would be up to you, but do not blame Maxent, you were the one who decided what the relevant knowledge was.

3. The Fallacy of Objectivity

A second objection that is, in my experience, raised against ME concerns examples such as the following:

Suppose I am asked what probability to give the next toss of a coin landing heads. Denote this event by p, so $\neg p$ corresponds to the coin landing tails instead. In this case I could take my (relevant) knowledge base K to be simply the empty set. Doing this we get, as would be expected, $ME(\emptyset)(p) = 1/2$. (Perhaps one could come up with something beyond $\emptyset$ but in any case the underlying heads/tails symmetry here surely means that any knowledge about heads/tails should be duplicated by knowledge about tails/heads so for our purposes here adding it to K would not make any difference.)

At this point then I have come to believe that heads are at least as likely as tails, since I am giving them both the same probability $1/2$, and since it is something I believe I am surely permitted to add it to my knowledge base $K = \emptyset$ to produce the knowledge base

$$K' = \{\, w(\neg p) \leq w(p) \,\}, \text{ equivalently } K' = \{\, 1/2 \leq w(p) \leq 1 \,\},$$

without it affecting what I believe. For surely I should not be in the position of having to change my beliefs on receipt of something I already believed?! And indeed Maxent does not require me to do so, applying ME to this enhanced knowledge base K' gives $ME(K')(p) = 1/2$, same as before I incorporated this new belief. (As a referee pointed out we could similarly have taken the equivalent knowledge base $K'' = \{0 \leq w(p) \leq 1/2\}$, so the only reasonable probability should be in the intersection of these two intervals , *i.e.*, $w(p) = 1/2$.)

At this point, however, critics of Maxent claim this is now an unreasonable answer, that I believe $1/2 \leq w(p) \leq 1$ and hence a middling answer, that the probability of p is 3/4, is better than the answer of 1/2 which is right at the end of the range.

Within the approach to Maxent being taken as a quintessential expression of common sense, I believe the reason for this preference for 3/4 rather than 1/2 shows a fundamental lack of appreciation of what is going on here. The probabilities assigned by Maxent (in this context) are *subjective probabilities,* quantified expressions of degrees of belief. What they are not is *estimates of objective probabilities*. (There certainly are arguments for applying Maxent in the context of objective probabilities, for classic example(s) see [10,11], but they are not the concern of this paper and should not be allowed to confuse the issues.) The answer 3/4 might look OK if one was trying to minimize the difference

$$(\text{TrueProb}(p) - \text{EstimatedProb}(p))^2,$$

which is presumably the thinking behind those who advocate the answer 3/4, but Maxent is giving something else, an answer based on rational, or common sense considerations.

To appreciate this point further suppose that we were talking not about a toss of a coin but instead a two horse race, say horses B and W, about which I knew nothing, and I intended to turn my beliefs into a capital investment, in short a bet. Now there is no "true" probability so does 3/4 still look so reasonable given my enhanced beliefs K'? I think not. My new belief, that horse B is at least as likely to win as horse W, is completely consistent with the belief I had formed on the basis of no knowledge, in fact it is a logical consequence of what I already believed, so why should simply acknowledging it as "knowledge" then cause me to change my beliefs?

Naturally these examples raise other issues, for example does not it matter *how* I came to learn that heads was as likely as tails? Surely if someone told me that on 1000 previous tosses this coin had landed heads 750 times I could be lured away from my favored 1/2 probability. Well doubtless so, but we have now shifted from the probabilities being *my subjective* probabilities to being the coin's *objective* probabilities and even if these could somehow be reconciled my resulting knowledge would be considerably more than just the above K'.

4. Maxent and Insufficient Reason

In the above mentioned paper [7], and previously at [12] (p. 424), Seidenfeld appears to be thinking of Maxent as a souped up version of Laplace's Principle of Insufficient Reason—that if there is no reason for two sentences/events to have different probabilities then they should get the same probability (very reasonably he does not like Insufficient Reason, but by linking it with Maxent he attempts a hatchet job on both). However, what Maxent satisfies (in common with many other inference processes) is not the problematic Principle Insufficient Reason as most people understand it but rather a special version of it that *is* above criticism. (In the continuous, as opposed to the discrete context we are considering here, the Principle of Insufficient Reason is associated with a number of paradoxes, in particular von Mises *Water and Wine Paradox* and Bertrand's *Stick and Circle Paradox*. Interestingly in [13] Jaynes proposes a solution to this latter by considering essentially the sort of invariance under automorphisms that we detail here in the discrete case. Naturally this has not met with total acceptance and the debate continues, see for example [14–21].)

To explain this say that a bijection $\sigma : SL \to SL$ is an *automorphism* of SL if

$$\begin{array}{rcl} \theta \equiv \phi & \iff & \sigma(\theta) \equiv \sigma(\phi), \\ \sigma(\neg\theta) & \equiv & \neg\sigma(\theta) \\ \sigma(\theta \wedge \phi) & \equiv & \sigma(\theta) \wedge \sigma(\phi) \\ \sigma(\theta \vee \phi) & \equiv & \sigma(\theta) \vee \sigma(\phi) \\ \sigma(\theta \to \phi) & \equiv & \sigma(\theta) \to \sigma(\phi). \end{array}$$

Colloquially, $\sigma(\theta)$ is, up to logical equivalence, a copy or doppelganger, of θ. Given this faithful representation of θ as $\sigma(\theta)$ that such an automorphism provides we should expect that the relationship of $\sigma(\theta)$ to a knowledge base $\sigma(K)$ is the same as that between θ and K, where $\sigma(K)$ is the result of replacing each $w(\phi)$ occurring in K by $w(\sigma(\phi))$.

Given this relationship we might therefore feel it would be desirable if ME satisfied

$$ME(\sigma(K))(\sigma(\theta)) = ME(K)(\theta)$$

and this does indeed hold.

In many instances this looks like a version of "Insufficient Reason". For example if $q = 2$ and

$$K = \{w(p_1 \wedge p_2) = 1/3\}$$

then there is (believe me!) an automorphism σ of SL such that

$$\sigma(p_1 \wedge p_2) = p_1 \wedge p_2,$$

$$\sigma(p_1 \wedge \neg p_2) = \neg p_1 \wedge p_2,$$

$$\sigma(\neg p_1 \wedge p_2) = \neg p_1 \wedge \neg p_2,$$

$$\sigma(\neg p_1 \wedge \neg p_2) = p_1 \wedge \neg p_2.$$

and $\sigma(K) = K$.

For this automorphism we have that

$$ME(K)(p_1 \wedge \neg p_2) = ME(\sigma(K))(\sigma(p_1 \wedge \neg p_2)) = ME(K)(\neg p_1 \wedge p_2)$$

$$ME(K)(\neg p_1 \wedge p_2) = ME(\sigma(K))(\sigma(\neg p_1 \wedge p_2)) = ME(K)(\neg p_1 \wedge \neg p_2)$$

$$ME(K)(\neg p_1 \wedge \neg p_2) = ME(\sigma(K))(\sigma(\neg p_1 \wedge \neg p_2)) = ME(K)(p_1 \wedge \neg p_2)$$

and this forces

$$ME(K)(p_1 \wedge \neg p_2) = ME(K)(\neg p_1 \wedge p_2) = ME(K)(\neg p_1 \wedge \neg p_2)\ (= 2/9)$$

—just like "Insufficient Reason".

In this case we are justified in using "Insufficient Reason" by dint of there existing an automorphism, or symmetry, of SL under which the knowledge is invariant (again there are many other inference processes which satisfy this same property, it is certainly not special to Maxent). As far as Maxent is concerned this is what you get, and it is beyond reproach. To my knowledge there is no mythical higher principle of "Insufficient Reason" to which Maxent blindly defers.

5. Conclusions

Most of us would surely prefer modes of reasoning which we could follow blindly without being required to make much effort, ideally no effort at all. Unfortunately, Maxent is not such a paradigm; it requires us to understand the assumptions on which it is predicated and be constantly mindful of abusing them.

Conflicts of Interest

The author declares no conflict of interest.

References

1. Paris, J.B. Common Sense and Maximum Entropy. *Synthese* **1999**, *117*, 75–93.
2. Paris, J.B.; Vencovská, A. A Note on the Inevitability of Maximum Entropy. *Int. J. Approx. Reason.* **1990**, *4*, 183–224.
3. Shore, J.E.; Johnson, R.W. Axiomatic Derivation of the Principle of Maximum Entropy. *IEEE Trans. Inf. Theory* **1980**, *26*, 26–35.
4. Csiszár, I. Why Least Squares and Maximum Entropy? An axiomatic approach to inference for linear inverse problems. *Ann. Stat.* **1991**, *19*, 2032–2066.
5. Paris, J.B.; Vencovská, A. Common Sense and Stochastic Independence. In *Foundations of Bayesianism*; Corfield, D., Williamson, J., Eds.; Kluwer Academic Press: Dordrecht, The Netherlands, 2001; pp. 203–240.
6. Paris, J.B.; Vencovská, A. In defense of the Maximum Entropy Inference Process. *Int. J. Approx. Reason.* **1997**, *17*, 77–103.
7. Seidenfeld, T. Entropy and Uncertainty. In *Foundations of Statistical Evidence*; MacNeill, I.B., Humphrey, G.J., Eds.; Reidel: Boston, MA, USA, 1987; pp. 259–287.

8. Grünwald, P. Maximum Entropy and the glasses you are looking through. In Proceedings of the Sixteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI 2000), Stanford, CA, USA, 30 June–3 July 2000; pp. 238–246.
9. Halpern, J.Y.; Koller, D. Representation dependence in probabilistic inference. *J. Artif. Intell. Res.* **2004**, *21*, 319–356.
10. Jaynes, E.T. Information Theory and Statistical Mechanics. *Phys. Rev.* **1957**, *106*, 620–630.
11. Jaynes, E.T. Information Theory and Statistical Mechanics. II. *Phys. Rev.* **1957**, *108*, 171–190.
12. Seidenfeld, T. Why I am not an Objective Bayesian; Some reflections prompted by Rosenkrantz. *Theory Decis.* **1979**, *11*, 413–440.
13. Jaynes, E.T. The Well-Posed Problem. *Found. Phys.* **1973**, *3*, 477–493.
14. Aerts, D.; de Bianci, M. Solving the Hard Problem of Bertrand's Paradox. *J. Math. Phys.* **2014**, *55*, doi:10.1063/1.4890291.
15. Bangu, S. On Bertrand's Paradox. *Analysis* **2010**, *70*, 30–35.
16. Gyenis, Z.; Rédei, M. Defusing Bertrand's Paradox. *Br. J. Philos. Sci.* **2014**, doi:10.1093/bjps/axt036.
17. Klyve, D. In defense of Bertrand: The non-restrictiveness of reasoning by example. *Philos. Math.* **2013**, *21,* 365–370.
18. Marinoff, L. A resolution of Bertrand's Paradox. *Philos. Sci.* **1994**, *61*, 1–24.
19. Rowbottom, D. Bertrand's Paradox revisited: Why Bertrand's "solutions" are all inapplicable. *Philos. Math.* **2013**, *21*, 110–114.
20. Rowbottom, D.; Shackel, N. Bangu's random thoughts on Bertrand's Paradox. *Analysis* **2010**, *70*, 689–692.
21. Shackel, N. Bertrand's Paradox and the Principle of Indifference. *Philos. Sci.* **2007**, *74*, 150–175.