A Note on Foundations of Bayesianism

Stefan Arnborg* stefan@nada.kth.se
Kungl. Tekniska Högskolan, Nada,
SE-100 44 Stockholm, SWEDEN

Gunnar Sjödin sjodin@sics.se
Swedish Institute of Computer Science, Box 1263,
SE-164 29 Kista, SWEDEN

Abstract

We discuss the justifications of Bayesianism by Cox and Jaynes, and relate them to a recent critique by Halpern(JAIR, vol 10(1999), pp 67–85). We show that a problem with Halperns example is that a finite and natural refinement of the model leads to inconsistencies, and that the same is the case with every model in which rescalability to probability cannot be done. We also discuss other problems with the justifications and assumptions usually made on the function F describing plausibility of conjunction. We note that the commonly postulated monotonicity condition should be strengthened to strict monotonicity before Cox justification becomes convincing. On the other hand, we note that the commonly assumed regularity requirements on F (like continuity) or its domain (like denseness) are unnecessary.

1. Introduction

Several ways are possible for dealing with uncertainty and ignorance in AI applications and the correct way of dealing with it cannot be proved, since this is not a purely mathematical question. Bayesianism is one possible way that has been tried in so many application areas and with so convincing results that its proponents have claimed it superior to all its alternatives and for all applications. Such claims are easier to justify with some very fundamental scientific argument. One such argument could be that even if other ways to deal with uncertainty are possible, they either have some easily stated deficiency or are equivalent to Bayesianism.

Indeed, such arguments have been put forward and they have not been unanimously accepted. This also would follow from Bayesianism itself, since prior prejudices are predicted by the theory to outweigh every informal argumentation, and there is no proof method relating to real-world phenomena with the persuasiveness of pure logic and mathematics. This note was inspired by a recent critique (Halpern, 1999) of (Cox, 1946) and (Jaynes, 1996). We try to identify the separating point and propose a solution to the riddle posed by Halpern.

We try to find assumptions that are strong enough to strictly imply Bayesianism and at the same time convincing in a subjective way (common sense). In section 2 we give a short outline of Cox arguments and introduce the function F relating the plausibility of a conjunction to the plausibilities of its conjuncts. In section 3 we describe Halperns example. In section 4 we discuss the problem raised by the difference between the conclusions of the

*. SICS NNRC project

two papers. In subsection 4.1, we show that a simple and natural refinement of Halperns example leads to inconsistency, and that the same will be the case if the function F has a violation of associativity or symmetry. In 4.2 we observe that the difference between probability and possibility is the insistence on strict monotonicity instead of monotonicity of F. In 4.3 we show that even if there is no direct violation of strict monotonicity, associativity or symmetry, there can be problems in a model that surface after a number of refinement steps, and we describe a theorem (proved in the appendix) saying that for a finite domain, natural refinements are possible if and only if the plausibilities are rescalable to probabilities. In 4.4 we discuss the extension to infinite domains. We give conditions under which rescalability is possible. These conditions are related to a common sense description of desirable features. We do not find purely technical conditions such as continuity or some dense type of domain among these conditions. Theorems justifying our conclusions are proved in the appendix.

2. Arguments for the Inevitability of the Bayesian View

In 1946, R.T. Cox published his findings(Cox, 1946) on some properties required by any good calculus of plausibility of statements. He stated three requirements (the following is actually from Jaynes, but very similar):

I: Divisibility and comparability- The plausibility of a statement is a real number and is dependent on information we have related to the statement.

II: Consistency - If the plausibility of a statement can be derived in two ways, the two results must be equal.

III: Common sense - Aristotelian deductive logic should be the special case of reasoning only with statements known to be true or known to be false, and plausibilities should vary sensibly with the assessment of plausibilities of inputs.

The paper is very appealing to believers in Bayesianism, but sometimes more has been put in it than there is, and sometimes less. A very lucid elaboration of Cox findings can be found in Chapter 2 of E.T. Jaynes posthumous manuscript(Jaynes, 1996). After introducing the notation $A|C$ for the plausibility of statement A given that we know C to be true, he finds, using propositional logic, the governing functional equation for defining the plausibility of a conjunction: $AB|C = F(A|BC, B|C)$ must hold for some function F. Since $ABC|D \equiv (AB)C|D \equiv A(BC)|D$, F must satisfy the equation of associativity: $F(F(x, y), z) = F(x, F(y, z))$, with $x = A|BCD$, $y = B|CD$ and $z = C|D$. If this were true for all x, y, and z (this is the key element in Halperns critique), we could solve the equation after postulating some regularity condition on F. Solving the equation is non-trivial, but the result is that no matter what our choice of $A|C$ is, there must be a function w such that $w(AB|C) = w(A|C)w(B|AC)$. This follows, *e.g.*, by assuming F defined on a dense interval and twice continuously differentiable. By letting the constituent statements in the relation $w(AB|C) = w(A|C)w(B|AC)$ take on various combinations of truth and falsity, we find that we can, without loss of generality, demand that $w(A)$ goes from 0 when A is falsity, monotonically to 1 when A is truth.

There must also be a function S with the property $S(w(A|C)) = w(\overline{A}|C)$. By manipulating this equation with the rules of propositional logic we find that S must satisfy

$xS(S(y)/x) = yS(S(x)/y)$ for $y \geq S(x)$. This is also a fairly difficult functional equation for S, but it is relatively easy to verify that one family of solutions is given by $S(x)^m + x^m = 1$, for real number $m > 0$. These are also the only solutions. Consequently, every type of reasoning with the plausibility of statements satisfying I, II and III above is equivalent to computing with probabilities after the rescaling of the plausibility measure to $w(x)^m$. The two laws we found are (switching now to the familiar notation P for probability) $P(AB|C) = P(A|C)P(B|AC)$ and $P(\overline{A}|C) = 1 - P(A|C)$. Bayes rule $P(A|BC) = P(B|AC)P(A|C)/P(B|C)$ is an immediate consequence of the probability rule for conjunction and commutativity of conjunction.

From this all the machinery of Bayesian analysis follows, except the way to assign prior probabilities. Thus, Bayesian analysis is equivalent to every way of reasoning with plausibility satisfying the criteria I, II and III, as interpreted by Cox and Jaynes. A similar derivation (more related to de Finettis work, so the probabilities are derived from consistent gambling behaviour) is (Lindley, 1982). The accompanying discussion is recommended reading as an illustration of how difficult this topic is. It must be said that neither Cox nor Jaynes are completely rigorous in defining their assumptions, and a recent critique can be found in (Halpern, 1999). It is clear that Halpern and Cox/Jaynes interpret condition III differently, since Cox solves a differential equation whereas Halpern gives a discrete counterexample. In his example the function F is not associative and whence his plausibility measure cannot be rescaled to probabilities, because non-associativity must be preserved by scaling. Cox/Jaynes do not state theorems in their texts. Therefore, Halperns paper cannot give a counterexample to a theorem by Cox, as the title of Halperns paper suggests. But it can be taken as evidence that Cox common sense assumptions are not the only ones possible. We do not think anyone would argue against the desirability of the three conditions I, II and III as given above. However, they may conflict with other desirable conditions. In particular, common sense is a rather open-ended condition and it can certainly be debated what is required by common sense, and what is not.

The first condition I has been characterized as the 'dogma of precision' and is sometimes found unacceptable essentially by arguments saying that we cannot know which exact real numbers to use. Several alternatives based on intervals instead of numbers have been designed and motivated, by see, *e.g.*, (Walley, 1996) and others. This may in turn lead to problems in choosing the exact real values used as end-points of the intervals. Although some interval based schemes can be seen as multiple-context Bayesian inference in analogy with multiple-criteria decision making, they are usually not presented as such. There are many possible objections, but here we will concentrate on a set of problems discussed by Halpern. We will not give a full account of the background to the discussions and developments of Cox ideas. There is a fairly comprehensive discussion in (Halpern, 1999). For more far-reaching discussions, see, *e.g.*, (Jaynes, 1996), (Walley, 1991) and (Fine, 1973). Paris gives a full-fledged proof of rescalability (Paris, 1994) that does not use differentiability assumptions, but insists on an often omitted density assumption of the set of probability values (which cannot hold on a finite setting).

Many approaches were analysed in the imprecise probabilities project(Walley, 1999). Particularly, there are many ways to resolve the difficult distinction between ignorance and uncertainty by making hierarchical models involving both possibility and probability. This is not a central theme of this note, however.

3. Halperns Example

It only follows from the argument of Cox/Jaynes (referring to condition III) that associativity holds for values actually occurring as plausibilities of statements $A|BCD$, $B|CD$ and $C|D$. Halpern designs a small world where there are no 4-tuples of statements to which the associativity condition could apply. In the notation of Cox, the example consists of four groups of three statements each($\rightarrow$ stands for implication):

A, B, C, where $A \rightarrow B$ and $B \rightarrow C$ hold
D, E, G, where $D \rightarrow E$ and $E \rightarrow G$ hold
H, I, J, where $H \rightarrow I$ and $I \rightarrow J$ hold
K, L, M, where $K \rightarrow L$ and $L \rightarrow M$ hold

C, G, J and M exclude each other, so only one of them can hold. Plausibilities are assigned to these statements as follows:

$D|EG = H|IJ = 3/5$
$E|G = A|BC = 5/11$
$B|C = L|M = 11/19$
$A|C = I|J = 5/19$
$D|G = K|LM = 3/11$
$K|M = 3/19$
$H|J = 3/19 - \delta$, for some small $\delta > 0$.

From the above we find $5/19 = A|C = AB|C = F(A|BC, B|C) = F(5/11, 11/19)$ and $3/11 = D|G = DE|G = F(D|EG, E|G) = F(3/5, 5/11)$. Moreover, $3/19 = K|M = KL|M = F(K|LM, L|M) = F(3/11, 11/19)$, but $3/19 - \delta = H|J = HI|J = F(H|IJ, I|J) = F(3/5, 5/19)$. It is easy to see that these plausibilities are consistent in all ways, as shown by the detailed model in (Halpern, 1999). With the exception of $H|J$, all quantities could have been probabilities. The function F is not associative, because $F(3/11, 11/19) = F(F(3/5, 5/11), 11/19) = 3/19$, but $F(3/5, 5/19) = F(3/5, F(5/11, 11/19) = 3/19 - \delta$.

4. Discussion

What happened in the example is that the same plausibility values were assigned to seemingly unrelated conditional statements. Therefore, a violation of associativity yields no immediate inconsistency. It appears quite clear that there is a difference in the assumptions made by Cox and those made by Halpern: Cox assumes implicitly that F is a universal function, to be used for all possible models. This is not entirely unreasonable if the analogy between F and the truth table for conjunction is seen - we seldom let propositional connective truth tables be model dependent. Halpern on the other hand - also implicitly - assumes that every model could have its own function F. This is also quite reasonable, and there is for example the even more relaxed condition that several F functions can be used in the same model. But we do not want to leave this problem without a somewhat deeper analysis. Instead of discussing contrived technical conditions under which Bayesianism can be recovered, we argue that a well-informed method choice can be obtained in a dialogue around desirable properties of an uncertainty measure and methodology. Such a dialogue can focus on questions like these, after a general consensus about conditions I and II of Jaynes:

- **Refinability:** Assume we have a consistent model with some statement S. Should it be possible to refine S into two parts S' and S'', so that any value already appearing in the model could be given to the plausibility $S'|S$?

- **Strict Monotonicity:** Will it always be the case that the plausibility of a conjunction is less than those of the conjuncts, if these are independent and their plausibilities are not 0 or 1?

- **Separability:** Is there a 'separating plausibility' c in the model with the following property: Consider any two plausibilities x and y of the model, $x < y < 1$. Consider three refinement sequences X_i, Y_i and C_i such that the plausibility of $X_{i+1}|X_i$ is constant x, $Y_{i+1}|Y_i$ is constant y, and $C_{i+1}|C_i$ is constant c. Are there then always integers p and q such that $X_p|X_0 < C_q|C_0 \leq Y_p|Y_0$?

This dialogue might lead to discussions of what is meant by independence, representativity, ignorance and uncertainty, and also of which the space is to which plausibilities are assigned. But the questions above are relevant. We mean that 'yes' answers to them are minimal precise conditions that entail Bayesianism. No purely technical conditions, introduced just to make the proof go through, are necessary.

4.1 Refinability

Halpern takes his example as an indication that the 'proof' of Cox only applies to infinite domains with a particular type of (dense) plausibility assignment, and observes that such cases are unusual (indeed non-existent) in practice. There are, however, more conditions that can be extracted from the desideratum of common sense. One such can be called refinability - it can be claimed reasonable that certain types of refinements of the model should be possible, and not depend a lot on somewhat arbitrary details designed into other parts of the model. In the example, it would be perfectly reasonable that one wants to add a new statement A' in the C part of the model, and such that $A' \to A$. Moreover, it would be reasonable to allow any value to be assessed for the plausibility $A'|A$, because there is no link to the rest of the model. But there is one value, namely 3/5, that cannot be assigned, because this would give a violation of the associativity law for the statement $A'AB|C$, which could be proved to have plausibility both 3/19 and $3/19 - \delta$. One can claim that this effect is a violation of common sense. It involves nothing that is infinite. It is readily seen that it is generally valid: if a model has a violation of associativity, then there exists a simple and finite refinement (in three steps) that is arbitrarily blocked. A similar argument can be used to show that F is bound by common sense to be symmetric: It is reasonable that one should be able to add two conditionally independent statements (by independence we mean, of course, that $A|BC = A|C$ and $B|AC = B|C$) with any plausibilities to the model, and then symmetry of F follows from commutativity of conjunction. Symmetry was never assumed by Cox, but followed from the differentiability properties assumed for F. The above is meant to suggest that it is quite reasonable to require that the function F is associative and symmetric for all values that appear in a given model. These are the properties postulated for the similar functions T-norms used in (Bonissone & Wood, 1989) and by many others. These functions were also postulated to be monotonic, however.

4.2 Strict Monotonicity

Another point raised by Halpern is the status of Dubois-Prades possibilistic system where F is proposed to be the minimum function. There are important uses of this function(Benferhat, Dubois, & Prade, 1997), but not as the function F in Bayesianism. One must ask if there is a common sense reason to exclude it rather than finding various contrived technical conditions that block it. The min function arises naturally from a common sense observation: it is reasonable to assume that $AB|C = F(A|BC, B|C) \leq \min(A|BC, B|C)$. The min function is thus an upper bound for possible F functions. Cox states that he assumes F to be twice continuously differentiable. But this is just a contrived condition required for his proof method. It seems unreasonable to exclude the min function for the reason that it is not twice continuously differentiable, or at least to argue that common sense prescribes differentiability. It is somewhat easier to argue that F must be strictly monotone when none of its arguments represents falsity (*i.e.*, if x is not falsity and $u > v$, then $F(u, x) > F(v, x)$ and $F(x, u) > F(x, v)$). Jaynes states the requirement of strict monotonicity, although he does not mention when he uses it: "If $A|C$ becomes more plausible, and $B|AC$ is not falsity, then $AB|C$ also becomes more plausible, if nothing else (namely $B|AC$) changes". This statement can certainly not be verified mathematically, but is something you have to believe to accept. But it tallies well with numerous observations describing the relationship between uncertainty and ignorance(Wakker, 1999), and seems closely related to the additivity assumption used in de Finettis and Lindleys framework.

If possibility is related to ignorance, as many claim, ought one not be able to use a function F that is not strictly monotone, like the min function? This is certainly not excluded. But in practice it seems as if two-stage approaches are very successful for problems with both ignorance and uncertainty. Thus, the standard Bayesian hierarchical models(Carlin & Louis, 1997) can be described as probability distributions over probability distributions. Robust Bayesian methods use sets of probability distributions(Berger, 1994), and the sets of desirable gambles(Walley, 1999) are quite similar and considered a candidate for a most general framework of imprecise probability. Belief functions are probabilities over possibilities (or, more precisely, probabilities of sets(Wakker, 1999)).

4.3 Completion

A last question raised by Halpern is whether any partially specified function can be extended to an associative function if it is associative on its range of definition. This is not generally the case, even if it also satisfies the other properties that will be required from the completed function: strict monotonicity and symmetry. The rescaling operation defined by Cox involves an arbitrary (non-negative) function $H(x)$:

$$w(x) \equiv \exp(\int^{x} \frac{dx}{H(x)})$$

that he shows to exist for all associative and twice continuously differentiable functions F (we may have to add strict monotonicity or improve the proof). If we take this function H to be piecewise constant between the points appearing in the model, we get an extension to an associative and piecewise twice continuously differentiable function by solving a finite linear system of equations and inequalities for the quantities $l_i = \log w(x_i)$ excluding the value for falsity. The system has an equation $l_i + l_j = l_k$ for each triple $x_k = F(x_i, x_j)$ and an inequality $l_i < l_j$ for every pair with $x_i < x_j$, and an equality $l_i = l_j$ when $x_i = x_j$.

If a partially specified function can be completed to a full function over the support points (and some more points) satisfying associativity, symmetry and strict monotonicity, then the system is solvable. Its solution set is either unbounded or empty, and it is empty only if there is no completion satisfying associativity, symmetry and strict monotonicity. Let us give a slightly more complex example showing how this works – the formal part is given in Theorem 3 of the appendix. A simple case where the partially specified function triples satisfy the laws, but no completion over the support points does so, is the following: Assume the partial specification satisfies

$$F(x_4, x_4) = F(x_3, x_5) = a \tag{1}$$
$$F(x_2, x_4) = F(x_1, x_5) = b \tag{2}$$
$$F(x_4, x_6) = F(x_3, x_7) = c \tag{3}$$
$$F(x_2, x_6) = F(x_1, x_8) = d \tag{4}$$

Here we have assumed that the x_i quantities are ordered increasingly in the open interval $(0, 1)$, but the quantities a, b, c and d can have any values. If the plausibilities were scalable to log probabilities l_i, there should be a solution to the system:

$$l_4 + l_4 = l_3 + l_5 \tag{5}$$
$$l_2 + l_4 = l_1 + l_5 \tag{6}$$
$$l_4 + l_6 = l_3 + l_7 \tag{7}$$
$$l_2 + l_6 = l_1 + l_8, \tag{8}$$

together with the conditions $l_i < l_{i+1}$. We would be able to decide from (5) and (6) that $l_4 - l_2 = l_3 - l_1$ and thus $F(x_1, x_4) = F(x_3, x_2)$ by simple elimination, and we could similarly work out the consequence $l_7 = l_8$, contrary to the condition $l_7 < l_8$. But if it were possible to complete the partially specified F so that it satisfies symmetry and associativity, we can reach the same conclusion by observing that the first two equations can be combined with F to yield: $F(F(x_4, x_4), F(x_1, x_5)) = F(F(x_3, x_5), F(x_2, x_4))$ and thus by symmetry and associativity $F(F(x_4, x_5), F(x_1, x_4)) = F(F(x_4, x_5), F(x_2, x_3))$, which entails, because of strict monotonicity and $0 < F(x_4, x_5) < 1$, that we have $F(x_1, x_4) = F(x_2, x_3)$. Continuing in the same fashion we obtain from (1) and (3): $F(x_4, x_7) = F(x_5, x_6)$, from (2) and (4): $F(x_4, x_8) = F(x_5, x_6)$, and finally $F(x_4, x_7) = F(x_4, x_8)$ and $x_7 = x_8$, contrary to the assumption that $x_7 < x_8$. This also means that it is possible to add a finite set of independent events with their plausibilities to the model that leads to inconsistency in the plausibility assignment. In this example we can add statements $\{A_i\}_{i=1}^{7}$, B_4 and C, with $A_i|C = x_i$ and $B_4|C = x_4$. If the A_i and B_4 are independent given C, the statement $A_1A_2A_3A_4B_4A_5A_6A_7|C$ can be shown to have two different plausibilities, $F(q, x_7)$ and $F(q, x_8)$ for $q = F(x_1, F(x_2, F(x_3, F(x_4, F(x_4, F(x_5, x_6))))))$. Theorem 3 shows that all finite refinement sequences of a consistent finite model are consistent if and only if the plausibility measure is rescalable to probability. We need not consider arbitrarily long sequences, since the length of a shortest inconsistent sequence can be bounded in terms of the size of the original model.

The above means that we can assign plausibilities in two different ways: either we choose a function F that has the required properties and use it for assigning plausibilities of conjunctions, or else we assign plausibilities on the fly but check always (by solving an LP problem) that no newly defined triple (arguments and function value) violates the required properties. In both cases it would be better to work with probabilities.

4.4 Separability

We finish the discussion by noting that Aczél, in another of his numerous books, has stated weaker conditions than twice continuous differentiability under which the solutions F to the equation of associativity can be expressed by $l(F(x, y)) = l(x) + l(y)$ (think of l as the logarithm of w). These conditions are continuity and cancelability (Kürzbarkeit): if $F(u, x) = F(v, x)$ or $F(x, u) = F(x, v)$, and $x \neq 0$, then it must be the case that $u = v$ ((Aczél, 1961), section 6.2). Cancelability and continuity leads immediately to strict monotonicity, as well as strict monotonicity independently of continuity entails cancelability. For completeness, the proof that our interpretation of common sense, with a universal function F with dense domain $[0, 1]$, leads to Bayesianism, is given as Theorem 4 in the appendix. This proof is a specialized version of the proofs in the work of Aczél on functional equations. This makes it easier to read and one will not have to consider the many special cases that can occur in general. Several authors have referred to these proofs as long and complex. This is unfortunate, because what is required for the Bayesian connection is easy.

The replacement of continuity and associativity assumptions by strict monotonicity and refinability entails Bayesianism in finite domains. Now it remains to consider non-finite domains. There seem to be no principled reason that Theorem 3 should not work in infinite domains. However we have not analyzed this problem, and in particular we do not think that finite refinability is sufficient for infinite domains. Instead we solve a slightly easier problem: Suppose that a model is defined, and its F function is completed to a minimal function that already covers all refinements. Which are the properties required for rescalability of such a function? If the domain and range of F is D and R, respectively, and $R \subset D$, then we need only one new condition before we can prove rescalability(Theorem 5 of the appendix), and this is the Separability condition of section 4! The Separability condition is only relevant for infinite domains, and there it has the consequence that there must be an accumulation point in the domain at the low end (typically 0). The Separability condition seems related to the Archimedean order axiom of (Fine, 1973), in the sense that it binds the domain together. One can certainly question the common sense inevitability of separability. But it is an easily stated condition that is much weaker but more robust than continuity and the accompanying domain denseness, which seem to us rather fragile conditions.

5. Conclusions

The Bayesian method of dealing with uncertainty is an important achievement of 20th century philosophy of science. It is based on very fundamental and inescapable principles. But its supremacy cannot be proved, only made plausible. In particular, Cox' argument is not in itself an argument for the three conditions of numerical plausibility, consistency and common sense.

We proposed to weaken Halperns proposed restriction for deriving Cox' result from infinite domains to the requirement of finite refinability. This implies that the triples for F used in a finite model must satisfy the requirements of associativity, symmetry and strict monotonicity, and it should be possible to extend the model with any finite set of statements that are independent and have the same plausibilities as those of the original model. The requirement of strict monotonicity is suggested by common sense, although it is in no way inevitable. The same holds for our refinability requirement, but it seems rather artificial to ignore it and one would like to see a plausible reason to drop it before doing so.

In an infinite model, the refinability condition means that a function F must be used which is associative, symmetric and strictly monotone. If we add the condition that the plausibility of an iterated conjunction approaches 0 if the conjuncts are of equal plausibility and independent, we can prove that Bayesianism rules.

As an interesting continuation of this line of inquiry, one can investigate what the exact consequences would be of replacing the strict monotonicity assumption with just monotonicity.

We would finally like to thank the referees for helpful comments.

References

Aczél, J. (1961). *Vorlesungen über funktionalgleichungen und ihre anwendungen.* Birkhäuser Verlag, Basel.

Benferhat, S., Dubois, D., & Prade, H. (1997). Nonmonotonic reasoning, conditional objects and possibility theory. *Artificial Intelligence*, *92*, 259–276.

Berger, J. O. (1994). An overview of robust Bayesian analysis (with discussion). *Test*, *3*, 5–124.

Bonissone, P., & Wood, N. C. (1989). T-norm based reasoning in situation assessment applications. In Kanal, L. N., Levitt, T., & Lemmer, J. (Eds.), *Uncertainty in Artificial Intelligence*, pp. 241–256. Elsevier Science Publishers.

Carlin, B. P., & Louis, T. A. (1997). *Bayes and empirical Bayes methods for data analysis.* Chapman & Hall.

Cox, R. (1946). Probability, frequency, and reasonable expectation. *Am. Jour. Phys.*, *14*, 1–13.

Fine, T. L. (1973). *Theories of Probability.* Academic Press.

Halpern, J. (1999). A counterexample to theorems of Cox and Fine. *Journal of AI research*, *10*, 67–85.

Jaynes, E. T. (1996). *Probability Theory: The Logic of Science.* Preprint: Washington University. `ftp://bayes.wustl.edu/Jaynes.book/`.

Kuhn, H. W. (1956). Dual systems of homogeneous linear relations. In Kuhn, H. W., & Tucker, A. W. (Eds.), *Linear inequalities and related systems*, pp. 3–18. Princeton University Press.

Lindley, D. V. (1982). Scoring rules and the inevitability of probability (with discussion). *Internat. Stat. Rev*, *50*, 1–26.

Paris, J. (1994). *The Uncertain Reasoners Companion.* Cambridge University Press.

Wakker, P. (1999). Dempster-belief functions are based on the principle of complete ignorance. In de Cooman, G., Cozman, F. G., Moral, S., & Walley, P. (Eds.), *Proceedings of the First International Symposium on Imprecise Probabilities and their Applicaations.* Gent University.

Walley, P. (1991). *Statistical Reasoning with imprecise probability.* Chapman and Hall.

Walley, P. (1996). Measures of uncertainty in expert systems. *Artificial Intelligence*, *83*, 1–58.

Walley, P. (1999). Towards a unified theory of imprecise probability. In de Cooman, G., Cozman, F. G., Moral, S., & Walley, P. (Eds.), *Proceedings of the First International Symposium on Imprecise Probabilities and their Applicaations.* Gent University.

Appendix A. Rescalability Theorems

We now return to the discrete case and show how a partially specified function F can be completed to a function allowing rescaling to Bayesian probabilities. Although one would maybe expect this to be a rather easy consequence of linear algebra, it turned out not to be quite as easy in its full generality. Halpern gives an example of a model that is not associative even on the original support set, but our example (section 4.3) is an example where non-extensibility to some finite extension base follows from the following theorem. More complex examples can easily be designed, once the mechanisms are grasped. First some definitions: We use the infix operator notation $x \circ y$ for $F(x, y)$, which is convenient for associative and symmetric functions. An *extension base* B of a sequence X of length L is a sequence (n_i) of length L of non-negative integers. A partial function that is associative and symmetric on X^2, where $X = (x_i)$, can be *extended to extension base* B if it can be extended to an associative and symmetric function on a domain such that the expression $v_1 \circ v_2 \circ \cdots \circ v_n$ and all its subexpressions have values if every v_i is equal to some x_j, and for all i, the number of occurences of x_i is not larger than the corresponding number n_i in B.

The following is a result in duality theory of linear programming((Kuhn, 1956), Corollary 1A, case (i)):

Lemma 1 (Kuhn) *The system of equations $Ax = 0$ has a positive solution $x > 0$ if there is no u such that $A^T u \geq 0$ and $u \neq 0$.*

We can now prove:

Lemma 2 *Let F be a linear subspace of R^n. The following conditions are equivalent:*

(i) There is no element in F with all components positive.

(ii) There is a nonzero vector d with non-negative components that is orthogonal to F.

Proof.

(ii)→(i) : This direction is obvious, since a vector orthogonal to a non-zero and non-negative one cannot have all components positive.

(i)→(ii) : Assume (i): There is no element in F of R^n with all components positive. Let F be the space spanned by the rows of matrix B, $F = \{B^T y : y \in R^k\}$. Let the rows of A be a base for the orthogonal co-space of F, $AB^T = 0$. Thus, $F = \{x : Ax = 0\}$ and $Ax = 0$ has no positive solution x by out assumption that (i) is the case. Since $Ax = 0$ has no positive solution, by Lemma 1 there is a u such that $A^T u \geq 0$ and $u \neq 0$. Now $u^T A$ is a non-negative vector, and it is orthogonal to every vector in F because $AB^T = 0$ and thus $(u^T A)(B^T x) = 0$ for all $x \in R^k$. So (ii) applies, *i.e.*, (i)→(ii).

Conditions (i) and (ii) are thus equivalent.
□

Theorem 3 *Let $X = (x_i)_{i=1}^L$ be an increasing sequence of distinct values in the open interval $(0,1)$. Let $S = \{1,\ldots,L\}$ and $T \subset S^3$ be a finite set of triples. The partial function $\circ$ satisfies $x_i \circ x_j = x_k$, for all $(i,j,k) \in T$, and $x_i \circ 1 = x_i$ for all $i \in S$.*
Then the following conditions (i) and (ii) are equivalent:

(i) There is a finite extension base B of X to which $\circ$ cannot be extended as a symmetric, associative and strictly monotone function.

(ii) There is no increasing sequence of positive numbers $(f_i)_{i=1}^L$ such that if $(i,j,k) \in T$, then $f_i + f_j = f_k$.

Proof.

(i)→(ii) If (ii) is not the case, there exists appropriate f_i. Define $l(x)$ by interpolation to a strictly monotone function between the constraints $l(x_i) = f_i$, $l(1) = 0$ and $\lim_{x\to 0} l(x) = \infty$. The function $x \circ y = l^{-1}(l(x) + l(y))$ is associative, symmetric and strictly monotone on $(0,1]$. So also (i) is not the case, which shows (i)→(ii).

(ii)→(i) Assume (ii) is the case. Define the $|T|$ by L matrix M to have one row for each tuple in T. For such a tuple (i,j,k), the row has the value 1 in columns i and j, the value -1 in column k, and zero otherwise. Matrix D is $L-1$ by L and has value $D = I' - I''$ where I' and I'' is the L by L unit matrix with the first and last row, respectively, deleted. Since (ii) is the case, there is no positive L-vector solution f to $Mf = 0$ that also satisfies $Df > 0$, since such a solution would contradict non-existence of the f_i.

The solution space F of $Mf = 0$ is such that the linear subspace DF is orthogonal to some non-zero vector d with non-negative components, by Lemma 2. In other words, a linear equation $d^T Df = 0$ for f can be derived from $Mf = 0$ only, $d^T D = c^T M$

for some vector c. Since M and D have integer coefficients, and the condition is homogeneous in d, we can assume that d consists of natural numbers and c of integers. Thus, a linear equality for f can be obtained as a linear combination with integer coefficients of the linear equalities encoded by the rows of M. But each row of M is derived from a constraint on the function $x \circ y$. By composing these constraints in the same pattern we can derive functional constraints on $x \circ y$, and at last obtain a functional constraint corresponding to the linear constraint $d^T Df = 0$. We compose the constraints coded by a triple of T a number of times given by the corresponding coefficient c_i of the linear combination, reversing the equation if the coefficient is negative. In this way we can derive functional constraints for $x \circ y$.

The linear constraint $d^T Df = 0$ can be written as $d_1 f_1 + d_2 f_2 + \cdots d_{L-1} f_{L-1} = d_1 f_2 + d_2 f_3 + \cdots d_{L-1} f_L$, where no d_i is negative and at least one is positive. This translates to $a_1 \circ a_2 \circ \cdots \circ a_m = b_1 \circ b_2 \circ \cdots \circ b_m$, where either $a_i = b_i$ or $a_i < b_i$, with at least one strict inequality since at least one d_i is non-zero.

But then, from strict monotonicity, we must also have: $a_1 \circ a_2 \circ \cdots \circ a_m < b_1 \circ b_2 \circ \cdots \circ b_m$.

There can thus not be a strictly monotone extension of $\circ$ to an extension base defined by the union of the (a_i) and (b_i) sequences, in other words (ii) is the case.

So (i) and (ii) are equivalent. □

An immediate corollary of Theorem 3 is that a model defined with a partial function $\circ$ satisfying the premises of Theorem 3, and which cannot be rescaled to a probability function, cannot consistently be refined by addition of a set of independent statements with plausibilities given by the quantities a_i and b_i whose existence were shown in the proof.

The following theorem is a simplified and tailored version of theorems due to Aczél(Aczél, 1961), which shows that Cox assumption that F is two times continuously differentiable can be replaced by continuity and strict monotonicity. The assumptions in this theorem are applicable if we regard $\circ$ as a universal function to be used in all models. The continuity and denseness assumptions will turn out to be technical conditions only needed because of the chosen proof method.

Theorem 4 *Let the function* $\circ : [0,1]^2 \to [0,1]$ *have the following properties:*

- *Associativity:* $(x \circ y) \circ z = x \circ (y \circ z)$ *for all* $x, y, z \in [0,1]$*;*
- *Strict monotonicity: if* $u < v$ *and* $z \neq 0$*, then* $u \circ z < v \circ z$ *and* $z \circ u < z \circ v$*;*
- *Continuity*
- $0 \circ x = x \circ 0 = 0$ *and* $1 \circ x = x \circ 1 = x$

Then $x \circ y = f(f^{-1}(x) + f^{-1}(y))$ *for an invertible function* f*.*

Proof. Define the function family $\phi_m(x)$ inductively by

- $\phi_0(x) = 1$

- $\phi_1(x) = x$
- $\phi_{m+1}(x) = x \circ \phi_m(x)$, for $m > 1$.

Abbreviate 'continuous and strictly monotone' by *csm.* By the associativity of $\circ$, we find $\phi_{m_1}(x) \circ \phi_{m_2}(x) = \phi_{m_1+m_2}(x)$ and $\phi_{m_1}(\phi_{m_2}(x)) = \phi_{m_1 m_2}(x)$. Since $\circ$ is csm, $\phi_m(x)$ is a csm function of x when $m \neq 0$, going from 0 to 1 as x does so. Therefore, we can define $\phi_{1/m}(x)$ as the unique solution y to the equation $\phi_m(y) = x$. This is also a csm function, with $\lim_{m\to\infty} \phi_{1/m}(x) = 1$. For every rational number $r = p/q$, the function ϕ_r is defined by $\phi_r(x) = \phi_p(\phi_{1/q}(x))$. This is a well-defined function since $\phi_{pn/(qn)}(x) = \phi_p(\phi_{1/q}(\phi_n(\phi_{1/n}(x)))) = \phi_{p/q}(x)$.

We now have $\phi_{r_1}(x) \circ \phi_{r_2}(x) = \phi_{r_1+r_2}(x)$ and $\phi_{r_1}(\phi_{r_2}(x)) = \phi_{r_1 r_2}(x)$. Moreover, if $x < 1$, then $\lim_{m\to\infty} \phi_m(x) = 0$, for if the limit is larger, say $a_x < 1$, then we would have $a_x = F(a_x, a_x) = F(1, a_x)$, violating strict monotonicity and continuity of $\circ$.

Select $c \in (0, 1)$ and let f be defined for rational number $r \geq 0$ by $f(r) = \phi_r(c)$. The function f is strictly decreasing on the rational numbers. For any rational number $r = p/q$ there can thus be a limit from below, $f(r-0)$, a limit from above, $f(r+0)$, and the function value $f(r)$. For a real number t, it can be defined as the limit of a non-decreasing sequence $r_i = p_i/q_i$ and a non-increasing sequence $r'_i = p'_i/q_i$, with $p_i + 1 = p'_i$. Then $f(r_i) \circ \phi_{1/q_i}(c)$ has limit $f(t-0) \circ 1 = f(t-0)$, by the continuity of F, but is also equal to $f(r'_i)$ which goes to $f(t+0)$. Thus, $f(t-0) = f(t+0)$ and hence f is continuous also for irrational number $t > 0$. Therefore, f can be uniquely extended to a csm function over the positive reals.

The function f obeys the law $f(r_1) \circ f(r_2) = f(r_1 + r_2)$, from which the theorem follows, since $f(r_1) \circ f(r_2) = \phi_{r_1}(c) \circ \phi_{r_2}(c) = \phi_{r_1+r_2}(c) = f(r_1 + r_2)$ for rational numbers, and by continuity also for real numbers r_1 and r_2. □

If the continuity restriction on $\circ$ is dropped, we cannot be sure that ϕ_m is continuous or that $\phi_{1/m}(x)$ exists, and thus f may be a partial function and f^{-1} discontinuous.

However, strict monotonicity is a strong condition on $\circ$, and it can only have an enumerable set of discontinuities on a path along which it is monotone. This is a standard result, and easy to verify by considering the list of discontinuites ordered by the jump magnitude. There can only be finitely many discontinuities for any jump magnitude, thus the whole list can be ordered. Indeed, we can prove that the continuity assumption, but not that of strict monotonicity, can essentially be dropped. By essentially we mean that we must introduce a symmetry assumption (motivated by common sense above) and the Separability assumption. We can see in the proof below that the Separability assumption appears truly required not only for the proof, but also for the theorem to hold, since it 'binds' the domain together. It may of course be replaceable by some simpler condition. It follows from the proof that one separating plausibility is enough to prove rescalability, and with rescalability every non-trivial plausibility is separating. It is also evident that if there is a non-separating plausibility, then the model is not rescalable. As an example violating the Separability condition and not being rescalable, consider a domain generated from three statements with plausibilities $C = 1/3$, $Y = 1/4$ and $X = 1/5$. Let $C^i \circ Y^j \circ X^k = 1/(1 + 3(i + j + k) + (j + 2k)/(i + j + k + 1))$. Now $X^p = 1/(3 * p + 2)$, $Y^p = 1/(3 * p + 1)$, and $C^q = 1/(3 * q)$, and separation is not obtained, beacuse no C^q

appears in any $[X^p, Y^p]$-interval. Dropping the integrality constraint on i, j, k makes the domain dense but the model inconsistent, so it is no counterexample to Theorem 4. An interesting observation on this model is that each of its finite subsets is rescalable.

The proof of Theorem 5 is quite simple, and even though we have not seen the result in the reference literature it may well be known in the research literature of functional equations. Recall our argument for associativity and symmetry: if one of these is violated, there is an inconsistent model obtainable by three refinement steps (with three new statements, each having a plausibility already occuring somewhere in the original model).

Theorem 5 *Let the function* $\circ : D^2 \to R$ *have the following properties:*

- $R \subset D$, $\{0,1\} \subset D$ *and* $D \subset [0,1]$;
- *Associativity;*
- *Strict monotonicity on* $D - \{0\}$;
- *Symmetry;*
- $0 \circ x = x \circ 0 = 0$ *and* $1 \circ x = x \circ 1 = x$
- *Separability: There is a* $c \in D$ *such that for all* $x, y \in D - \{0,1\}$ *and* $y < x$*, there are integers* p, q *such that* $y^p < c^q \leq x^p$.

Then for $x, y \in (D - \{0\})^2$, $x \circ y = f(f^{-1}(x) + f^{-1}(y))$*, for a partial strictly monotone function* f *whose inverse is a strictly monotone function* f^{-1}.

Proof. Define the function family $\phi_m(x)$ on D as before. It is well-defined since the range of $\circ$ is included in its domain. It is not necessarily continuous. As before, $\phi_{m_1}(x) \circ \phi_{m_2}(x)) = \phi_{m_1+m_2}(x)$ and $\phi_{m_1}(\phi_{m_2}(x)) = \phi_{m_1 m_2}(x)$, and $\phi_m(x)$ is a strictly monotone function of x when $m \neq 0$, going from 0 to 1 as x does so. The symmetry assumption also leads to the law $\phi_m(x \circ x') = \phi_m(x) \circ \phi_m(x')$. For a rational number $r = p/q$ and $x, y \in D$, we define $y = \phi_r(x)$ if $\phi_q(y) = \phi_p(x)$. Now, ϕ_r is a partial strictly monotone function. Let $f(r) = \phi_r(c)$ for some $c \in D - \{0,1\}$. As a function of r, $f(r) = \phi_r(c)$ is a strictly monotone decreasing partial function. We extend the inverse f^{-1} to real arguments by a direct limit process of interval bisection: For $x \in D - \{0,1\}$, define $f^{-1}(x)$ as the limit of p_i/q_i, when $q_i = 2^i$ and $\phi_{p_i+1}(c) < \phi_{q_i}(x) \leq \phi_{p_i}(c)$.

There is always a unique solution sequence p_i, because $\phi_0(c) = 1$, so the right inequality can be satisfied, and the left inequality can be satisfied because the separability condition makes $\lim_{n\to\infty} \phi_n(x)$ independent of x if $0 < x < 1$. The sequence p_i/q_i will converge to a real number, since the sequence p_j/q_j for $j > i$ is contained in the interval $[p_i/q_i, (p_i+1)/q_i]$, whose length goes to 0. Thus, we have defined $f^{-1} : (D - \{0\}) \to [0, \infty)$ as a strictly monotone function (strictly because of separability). Its range is closed under addition, as can be seen by comparing two bisection sequences $\{p_i\}$ and $\{p'_i\}$ converging to $f^{-1}(x) = r$ and $f^{-1}(y) = r'$ with the one, $\{p''_i\}$, for $d = x \circ y$ converging to $f^{-1}(x \circ y) = r''$. The relations $\phi_{p_i+1}(c) < \phi_{q_i}(x) \leq \phi_{p_i}(c)$ and $\phi_{p'_i+1}(c) < \phi_{q_i}(y) \leq \phi_{p'_i}(c)$ imply, because of symmetry of $\circ$, $\phi_{q_i}(d) \leq \phi_{p_i+p'_i}(c)$ and $\phi_{q_i}(d) > \phi_{p_i+p'_i+2}(c)$. This implies that $p''_i = p_i + p'_i$ or $p''_i = p_i + p'_i + 1$

for all i and thus $r'' = r + r'$. But then $f^{-1}(d) = f^{-1}(x \circ y) = f^{-1}(x) + f^{-1}(y)$, or $x \circ y = f(f^{-1}(x) + f^{-1}(y))$ for all $x, y \in (D - \{0\})$, which finishes the proof. $\Box$