Can the maximum entropy principle be explained as a consistency requirement?

Jos Uffink
Department of History and Foundations of Mathematics and Science
University of Utrecht
P.O. Box, 80.000, 3508 TA Utrecht
the Netherlands
e-mail: uffink@fys.ruu.nl

October 19, 1997

Contents

Abstract

The principle of maximum entropy is a general method to assign values to probability distributions on the basis of partial information. This principle, introduced by Jaynes in 1957, forms an extension of the classical principle of insufficient reason. It has been further generalized, both in mathematical formulation and in intended scope, into the principle of maximum relative entropy or of minimum information. It has been claimed that these principles are singled out as unique methods of statistical inference that agree with certain compelling consistency requirements. This paper reviews these consistency arguments and the surrounding controversy. It is shown that the uniqueness proofs are flawed, or rest on unreasonably strong assumptions. A more general class of

inference rules, maximizing the so-called Rényi entropies, is exhibited which also fulfill the reasonable part of the consistency assumptions.

1 Introduction

In any application of probability theory to the problems of science or practical life one meets the question of how to assess the probability of the occurrence of some event or of the truth of some hypothesis. And although the mathematical formalism of probability theory serves as a powerful probe when analyzing such problems, it cannot by itself answer this question. Indeed, the formalism necessarily remains silent on this issue, since its goal is just to provide theorems valid for all probability assignments allowed by its axioms. Hence, recourse is necessary to an additional rule which tells us in which case one ought to assign which values to probabilities. Such a rule must, of course, refer to the meaning of the concept of probability, and will hence be subject to debate and controversy.

In 1957 E.T. Jaynes proposed a rule to assign numerical values to probabilities in circumstances where certain partial information is available. Jaynes showed in particular how this rule, when applied to statistical mechanics, leads to the usual canonical distributions in an extremely simple fashion. He also showed that a general method of statistical inference could be built upon this rule, which subsumes the techniques of statistical mechanics as a mere special case. Today this rule, known as the maximum entropy principle (MEP), is used in many fields, ranging from physics and chemistry to image reconstruction and stock market analysis. The series of volumes on *Maximum Entropy and Bayesian Methods* and many other publications amply illustrates the wide interest in the subject.

Nevertheless the MEP has always remained controversial. In part this controversy derives from the fact that Jaynes' principle relies on a 'subjective' (also known as 'objective Bayesian' or 'neoclassical') interpretation of probability, as a measure of the degree of belief which a rational person ought to assign to the event. This contrasts with the ensemble or 'frequency of occurrence' interpretations which are more common in statistical physics. More generally, the controversy is related to the very goal of Jaynes' approach, namely to remove statistical mechanics from the field of physics and reconstruct it as a theory of statistical inference, i.e. as a branch of logic or epistemology. Jaynes (1978) presents a colourful personal recollection of the resistance which the MEP met from the physics community, especially because of this perspective. We shall not, however, go into this side of the debate. The works of Penrose (1979), Denbigh and Denbigh (1985), Lavis and Milligan (1985), Buck and Macaulay (1991), Balian (1991) and Dougherty (1993) provide insight in the pro's and con's of the MEP in relation to statistical physics. For our purpose it suffices to note that the term 'maximum entropy principle' as it is used in this paper is not a physical principle in the proper sense, and should be carefully distinguished from the 'entropy maximum principle' of Tisza (1966) and Callen (1960). The latter is not a rule of inference but a condition for thermodynamical equilibrium.

Jaynes' approach has not drawn objections only for its radical reconstruction of a traditional physical theory. Also authors more sympathetic to the field of statistical inference and subjective probability have raised serious criticism, in particular Shimony and coworkers (Friedman and Shimony, 1971; Dias and Shimony, 1981; Shimony, 1985), Seidenfeld (1979, 1986) and Van Fraassen e.a. (1981,1986). These critics argued that the MEP conflicts with other established rules of statistical inference, in particular with that of Bayesian condition-

alization. On the other hand, defenders of the MEP have claimed that the principle is the unique rule of statistical inference satisfying certain compelling 'consistency requirements'. This claim indeed appears already in Jaynes' first paper on the subject. It has been greatly generalized and elaborated in the work of Shore and Johnson (1981), Tikochinsky, Tishby and Levine (1984) and Skilling (1988, 1989). If true this claim should, of course, silence all criticism. After all, nobody would be eager to "claim the distinction of reasoning inconsistently", as Jaynes (1986) put it. The very fact that critics were not silenced, however, suggests that these uniqueness proofs do not entirely settle this issue. Van Fraassen, Hughes and Harman (1986) challenged the claim more explicitly by exhibiting two alternative rules which, they argue, are no less reasonable rules of inference.

It is the purpose of this paper to review and investigate the controversy surrounding the MEP as a rule for statistical inference. In particular, we shall examine three versions of the claim that the MEP is the unique consequence of consistency requirements. We shall argue that not all the requirements needed for a unique characterization of the MEP are in fact reasonable. It is shown that a slightly smaller set of reasonable requirements are fulfilled if and only if the rule belongs to a class of which the maximum entropy principle, as well as the alternative rules of Van Fraassen, Hughes and Harman are members. This is the class of rules to maximize a generalized entropy expression containing a free continuous parameter (the so-called Rényi entropies).

These results suggest that a fruitful generalization of the MEP is obtained by the class of maximum Rényi entropy principles, as a new 'continuum of inductive methods'. However, the question which entropy expression to maximize is not the only issue involved in the controversy. The truly weak spot of the MEP and the alternative rules envisaged here lies in the way constraints on probability distributions are formed from the given partial information. The procedure for constructing these constraints brings back many objections that the MEP is able to avoid at first sight. This will be argued in more detail in a sequel paper.

This paper is organized as follows. In section 2 we recall the objections which beset the classical precursor of the MEP, the principle of insufficient reason. Section 3 shows how the MEP succeeds in avoiding many of these objections. The remaining questions concern the extension of the rule to continuum problems and the justification of the choice for the entropy expression. The extension of the principle to the continuum is discussed in section 5. A well-known solution for this problem is obtained by replacing the entropy expression by the relative entropy. We emphasize the consequences of this replacement for the status and interpretation of the resulting maximum relative entropy principle (MREP). The remainder of the paper is devoted to the problem of justification. We analyze three approaches to the claim that the choice for the entropy expression is the consequence of consistency requirements in section 4, 6 and 7. Section 8 is devoted to a short comparison with the conditions formulated by van Fraassen, Hughes and Harman (1986). Section 9, finally, summarizes the conclusions.

2 The principle of insufficient reason

The maximum entropy principle was introduced by Jaynes as an extension of the principle of insufficient reason of Laplace. The controversy surrounding the MEP is also to a large extent inherited from the legacy of this notorious predecessor. It is therefore worthwhile

going back first to the classical principle of insufficient reason (PIR).

Consider a random variable x. The values of x may represent outcomes of an experiment, states of a physical system, or just label various propositions; they are what Laplace calls the possible 'cases'. We assume in this section that x can take on only a finite number of possible values: $x \in S = \{x_1, \ldots x_n\}$. Our problem is to assign probabilities to the various values of x. The PIR states: whenever we have no reason to believe that one case rather than any other is realized, or, as it is also put, in case all values of x are judged to be 'equally possible', then their probabilities are equal to each other, i.e.

$$p(x_i) = \frac{1}{n}, \quad i = 1, \ldots, n.$$

This principle thus relies on a symmetry in our belief or judgment in order to obtain numerical values for probabilities. The underlying motivation is, of course, that in this view the term probability should be understood as a degree of belief and hence, the uniform probability distribution represents exactly the situation where all possible states are equally credible. Laplace was not the first to make this connection between probability and belief. Earlier similar arguments can be found in the work of Leibniz (1678), James Bernoulli (1713) and Bayes (1763). Laplace was the first, however, to turn this rule into the cornerstone of a comprehensive theory of probability.

Since the middle of the last century, the principle of insufficient reason and its consequences have become subject to extensive criticism –and sometimes ridicule– by R.L. Ellis (1842), J.S. Mill (1843), G. Boole (1854), J. Venn (1866), J. von Kries (1871) and J. Bertrand (1889) to name the most prominent. In fact even the very name of the principle is bound to make one feel uneasy.[1] The critics could find easy ammunition in the liberal and sometimes naive usage Laplace, Poisson and others made of the principle, especially in applications to the probability of testimony and in the rule of succession. In this century, the combined authority of R.A. Fisher, R. von Mises, J. Neyman, E.S. Pearson and H. Reichenbach has discredited the principle even further. In fact, until its revival by Jaynes, the principle of insufficient reason had hardly any supporters at all, with the outstanding exceptions of J.M. Keynes and H. Jeffreys. It seems that the objections under which the principle of insufficient reason has succumbed can be divided into four types.

(i). The first objection concerns the underlying interpretation of the notion of probability. According to many present-day authors this notion represents or entails a statement about the relative frequency of occurrence of an event. To say that in a certain situation the probability distribution is uniform means, according to this view, that when this situation is realized many times, all possible cases occur about equally often. In the principle of insufficient reason, on the other hand, probability assignments are based on a symmetry in our judgment, i.e. on the absence of knowledge that would favour the occurrence of one case above the other. The objection is then that one cannot derive empirical predictions from a lack of knowledge. As Ellis put it clearly (1850): "Mere ignorance is no ground for any inference whatsoever. *Ex nihilo nihil.* It cannot be that because we are ignorant of the matter we know something about it."

[1] Laplace did not, as far as I know, name his principle, and it is not certain who invented the expression, clearly intended as a nickname. Many modern authors credit Von Kries (1871) for coining the phrase (as the *Princip des mangelnden Grundes*). Indeed Von Kries himself also claimed to be the originator of this term (see Von Kries, 1916). Yet Boole already refers to the 'principle of non-sufficient reason' in an essay of 1862 as if it were a common name.

Therefore, most authors who are sympathetic to the view that probability is an empirical notion reject the principle of insufficient reason, whereas those who accept the principle mostly maintain that probability should be seen as an epistemological notion: a probability distribution represents the state of knowledge or belief of a rational mind, with the uniform distribution corresponding to a state of ignorance about x.

(ii). The Bertrand paradox. The second type of objection is more technical. It was shown by examples of Von Kries, Bertrand and Von Mises that the PIR leads to paradox when it is applied to the case where x ranges over a continuum. Indeed, the obvious extension of the principle to such cases is to adopt a uniform probability density when we have no reason to believe in the realization of one possible value rather than any other. The problem is now that one can choose different parametrizations for a continuum, and a probability density that is uniform over x becomes non-uniform under a non-linear parameter transformation, say $y = x^3$. This conflicts with the intuition that in a state of ignorance our judgment ought to be invariant under reparametrization: if we are ignorant of x we are also ignorant of y. Similar problems are actually also encountered in the case of discrete variables. In this case too a mere difference in bookkeeping can lead to different probability assignments, as is shown in Bertrand's example of the golden and silver coins.

(iii). The third objection goes back to James Bernoulli's *Ars Conjectandi* (1713). Long before Laplace, Bernoulli was already quite familiar with the idea of assigning numerical values to probabilities based upon lack of information, and sometimes he is regarded as the originator of the PIR. But Bernoulli also argued that this idea was of very limited applicability. According to him, it could be used almost exclusively in games of chance. Outside of this restricted context, for example in judging the risk of death, it is often too difficult to specify, say, the number of possible diseases, let alone to judge whether they are equally possible or not. In those cases Bernoulli advocated another method, based on his famous law of large numbers.

(iv). The last famous objection we mention has been made by Reichenbach. He claimed that the PIR was circular on the grounds that the only sensible meaning one can give to the phrase 'equally possible' is, in fact, 'equally probable' (Reichenbach 1935, p. 339). This criticism is obviously unfair to Laplace, who explicitly clarified that he meant the term to refer to a judgment: "cas également possibles, c'est à dire tels que nous soyons également indécis sur leur existence" (Laplace 1829). But Reichenbach is right to draw attention to the vagueness in the notion of possibility. The analysis of Von Mises (1928) and Hacking (1971, 1975) shows that in common language the notion of possibility is even more ambiguous than the notion of probability itself, so that a principle that grounds probability assignments in judgments of possibility, if not circular, is still not very enlightening.

The objections listed above played an important role in the downfall and eventual (almost) universal abandonment of the PIR. I do not claim that the list is exhaustive or even exclusive. In fact it is not easy to give a fair discussion of the PIR since many authors differ greatly in their statement of the meaning of the principle. Modern texts like those of Jaynes (1957a) or Fine (1973) formulate the PIR as the requirement to assign the alternative cases equal probability "if there is no reason to think otherwise" or "in the absence of known reasons to the contrary". These formulations seem rather different from Laplace's own, in the sense that they make a judgment about the probability itself rather than about the occurrence of cases the criterion for a probability assignment. Thus they are much more vulnerable to the charge of circularity. There are also variations of the PIR such as the principle of cogent reason (the *Prinzip des zwingenden Grundes* of Czuber) which are also

not always stated clearly and hard to distinguish from the PIR.

The disreputable status of the PIR is best illustrated by quoting from Keynes, who in 1921 attempted to save some valid version of the principle from its many difficulties. He admitted that these difficulties were

> "responsible for the doubts which philosophers and many others have often felt regarding any practical application of the [probability] calculus. Many candid persons, when confronted with the results of probability, feel a strong sense of uncertainty of the logical basis upon which it seems to rest. It is difficult to find an intelligible account of the meaning of probability, or of how we are ever to determine the probability of any particular proposition; and yet treatises on the subject profess to arrive at complicated results of the greatest precision and most profound practical importance. The incautious methods and exaggerated claims of the school of Laplace have undoubtedly contributed towards the existence of these sentiments." (Keynes, 1973, p. 55)

Keynes proposed to relieve the principle from its bad reputation by renaming it the 'principle of indifference'. This name seems no improvement because it suggests unwanted connotations with the notion of preference. In a game of Russian roulette, for example, one may very well judge the location of the bullet in each of the chambers of a revolver as equally possible, without feeling indifferent on the matter.

3 The revival of insufficient reason by maximum entropy

The principle of maximum entropy is a generalization of the principle of insufficient reason. We start again from the assumption that the variable x can take values in a finite set $S = \{x_1, \ldots, x_n\}$. It is now assumed that some information about this variable is given which can be modeled as a constraint on the set of probability distributions over S. It is assumed that this constraint exhaustively specifies all relevant information about x. The principle of maximum entropy is then the prescription to choose that probability distribution p for which the Shannon entropy, i.e. the expression

$$H(p) = -\sum_i p(x_i) \log p(x_i) \tag{1}$$

is maximal under the given constraints.

The most simple and often studied type of constraint is the case where the expectation value of some function f has a given value:

$$\langle f \rangle = \sum_i f(x_i) p(x_i) = \alpha. \tag{2}$$

In that case a well-known argument using Lagrange multipliers shows that the probability distribution with maximum entropy is of the form

$$p_\beta(x) = \frac{e^{-\beta f(x)}}{Z(\beta)} \tag{3}$$

where the parameters β and Z are determined by the constraint and normalization conditions,

$$-\frac{d}{d\beta} \log Z(\beta) = \alpha \tag{4}$$

$$Z(\beta) = \sum_i e^{-\beta f(x_i)}. \tag{5}$$

The MEP contains the PIR as a special case. Indeed, in the absence of reasons, i.e. in the case where no or only trivial constraints are imposed on the probability distribution, its entropy $H(p)$ is maximal when all probabilities are equal. But then, as the 'son of insufficient reason', the principle of maximum entropy of course enherits all objections associated with its infamous predecessor. How does it cope with these?

(i). With respect to the objection that no empirical knowledge can be derived from ignorance, the MEP can only agree. Indeed, in the frequency interpretation of probability the MEP seems to make little sense at all. Therefore, Jaynes has often emphasized that in the present view probability is not meant to represent a factual property of the real world but rather a state of knowledge about the world. Probability theory is in this approach not an empirical science and one should not expect to derive empirical consequences from the MEP. The maximum entropy probability distribution only represents our best prediction or judgment based on the given information.

(ii). The Bertrand paradox. In 1973 Jaynes produced a powerful argument to resolve the Bertrand paradox in line with the Maximum Entropy method and showed how a satisfactory solution for this problem is obtained by consideration of the relevant symmetry group. However, in order to do so the principle has to be adapted so as to be applicable to a continuum. We shall discuss the technical changes necessary to obtain this extension in section 5. We shall also argue that these technical changes involve important conceptual changes which alleviate the alleged merely subjective aspect of the principle.

(iii). How does the MEP fare in relation to Bernoulli's objection? Again, it provides progress. Since the MEP allows for using partial information in the form of constraints, it has obviously a much wider applicability than the PIR. Of course this is not to say that all of Bernoulli's worries are solved. Not every case of partial information can be modeled as a constraint on probability distributions. Also, the question what to do when the number of possible cases is unknown remains as yet unsolved.

(iv). Does the MEP boil down to a circularity, like the PIR in Reichenbach's analysis? Obviously not, because the MEP is much more specific and general than its predecessor. The choice to maximize the Shannon entropy expression H is clearly not trivial. One can envisage many alternative rules different from the MEP that also generalize the PIR. For example, consider the rule to choose that probability distribution that maximizes

$$\tilde{H}(p) = \sum_i \phi(p(x_i)) \tag{6}$$

with ϕ a concave function. It can be shown (see Hardy, Littlewood and Pólya, 1934, p. 89) that in the absence of constraints, the maximum of this expression for all probability distributions with n possible events is again obtained when all probabilities are equal. Thus, a 'maximum $\tilde{H}$ principle' will also generalize the principle of insufficient reason. But in general, such a rule will lead to very different value assignments. Hence, one may ask why the choice $\phi(x) = -x \log x$ is singled out above other concave functions in (6). Thus by avoiding the threat of circularity one raises the problem of *justification* of the MEP.

To summarize, we can say that the Maximum Entropy Principle provides clarification or progress on all of the objections that proved fatal to the PIR. There is however also an urgent new problem, that of justification of the entropy expression. It is to this problem that we now turn.

4 Justification by consistency: the approach of Jaynes

The claim that the MEP is justified by an appeal to consistency already appears in Jaynes' original article (1957a). In this article Jaynes based his claim on the following theorem of Shannon (1948) [2]

Theorem 1 *If the expression* $H_n(p_1, \ldots, p_n)$ *for* $p_i \geq 0, \sum_i p_i = 1$ *and* $n \geq 2$ *satisfies the conditions:*

1. $H_2(p, 1-p)$ *is a continuous positive function of* p.
2. *For all* n, $H_n(p_1 \ldots, p_n)$ *is a symmetrical (i.e. permutation invariant) function of* $p_1, \ldots, p_n$.
3. *For all* $n \geq 2$,

$$H_n(p_1, \ldots, p_n) = H_{n-1}(p_1 + p_2, p_3, \ldots, p_n) + (p_1 + p_2) H_2\left(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}\right).$$

then H_n *has the form*

$$H_n = -K \sum_{i=1}^{n} p_i \log p_i$$

for some positive constant K.

This theorem shows that the entropy or information measure H (we shall drop the subscript n for simplicity) is uniquely singled out (up to a multiplicative constant) by the three assumptions above. Jaynes argued:

> "A[n] ...important reason for preferring the Shannon measure is that it is the only one which satisfies the condition of consistency represented by the composition law [i.e. the assumptions of Shannon's theorem]. Therefore one expects that deductions made from any other information measure, if carried far enough, will eventually lead to contradictions." (Jaynes, 1981, p. 9)

And again in 1963:

> "It has by now been amply demonstrated by many workers that the "information measure" introduced by Shannon has special properties of uniqueness and consistency that make it *the* correct measure of the "amount of uncertainty" in a probability distribution" (Jaynes 1981, p. 45)

The justification of the MEP is then that the maximum entropy distribution is *the* distribution that correctly corresponds with a maximal amount of uncertainty. It represents the only probability assignment that is "maximally noncommittal with regard to missing information", i.e. that, while obeying the constraint, does not assume any information which we actually do not have.

This type of argument clearly hinges essentially on the idea that the Shannon entropy is the only measure of uncertainty that complies with conditions of consistency. The sense in which the term consistency is meant in these quotations is not completely clear, however.

[2] Here the assumptions of the theorem are given in the version of Faddeev (1957), who gave the first rigorous proof of the theorem.

In logic, the term 'consistency' is used to refer to a theory which does not entail a contradiction. In this logical sense, the most one can understand by a 'condition of consistency' is the requirement to reject logical contradictions and to accept tautological truths. And although the quotation above refers to the avoidance of contradictions, it is quite obvious that the MEP, or indeed any inference rule whatsoever, cannot be derived from such a requirement alone. Thus, in the present quotations the term 'consistency' is not to be taken in this logical sense. What is meant, presumably, is that the assumptions of Shannon's theorem are intuitively appealing, or perhaps even rationally compelling conditions to be demanded of any measure of information.[3]

Jaynes often attributed this 'consistency' argument to Shannon. In 1985, he even stated that, were it not for this appeal to consistency, "The name 'Claude Shannon' or the term 'Information Theory' would be quite unknown today" (p. 135). In fact, however, Shannon did not appeal to consistency in the derivation of his theorem at all. On the contrary, he rather de-emphasized the importance of his uniqueness theorem by writing:

> "This theorem and the assumptions needed for its proof are *in no way necessary* for the present theory. It is given chiefly to lend a certain plausibility to some of our later definitions. The real justification of these definitions, however, will reside in their implications." (Shannon 1948, p.393, emphasis added)

Thus, Shannon's assumptions are used by Jaynes for a purpose which apparently is not his own. But perhaps the theorem can bear a stronger reading than Shannon himself argued for. So let us ask whether the assumptions on which the theorem rests can be seen as conditions of consistency, i.e. whether alternative information measures would be 'incorrect', 'lead to contradictions' or have other undesirable consequences.

In this respect, the assumptions of Shannon's theorem are not immune to criticism. To mention an obvious point first, the third assumption implicitly assumes a scale on which entropy or information is to be measured. Is that scale rationally compelling? Khinchin wrote on this issue:

> "...it is natural to express the amount of information ... by an increasing function of H. The choice of this function means the choice of some unit for the quantity of information and is therefore fundamentally a matter of indifference."(Khinchin 1957, p. 7.)

In particular, one may add, this choice should not matter if our only interest is in the question for which distribution H becomes maximal, and not in the value of that maximum. But the choice does matter in Shannon's assumptions since they characterize H up to a multiplicative constant, and not H^2 or $\exp H$, etc. Hence these assumptions, however natural or convenient they may be, involve also what is more appropriately called *convention*, rather than consistency requirement.

A more substantial drawback is that Shannon's assumptions cannot be extended to the continuous case in a straightforward way. As is well-known, and will be discussed in more

[3] In later articles (1968, 1973) Jaynes uses the term 'consistency desideratum' for the demand that "in two problems where we have the same state of knowledge we should assign the same subjective probabilities." To me this formulation seems to express the intended interpretation of probability rather than a demand of consistency.

detail in section 5, more mathematical ingredients are needed in order to obtain a meaningful measure of information for continuous probability distributions than are mentioned in Shannon's formulation.

The most serious objection against regarding the assumptions of Shannon's theorem as compelling concerns one of its properties which, also according to Shannon, should provide the real justification of the expression. To explain this property and the objection against it, it is necessary to consider the concept of *conditional* entropy. Consider two variables x and y. The entropies of x and y are given by

$$H(x) = -\sum_i p(x_i) \log p(x_i),$$

$$H(y) = -\sum_j p(y_j) \log p(y_j).$$

(N.B.: in the notation $H(x)$ x should not be thought of as the argument of a function, but as a mere label specifying the variable.) The conditional entropy of x given y is defined as

$$H(x|y) = -\sum_y p(y) \sum_x p(x|y) \log p(x|y) \tag{7}$$

and one can derive the relation :

$$H(x|y) \leq H(x) \tag{8}$$

where equality holds just in case the variables x and y are independent. (To prove (8), use Jensen's inequality for the concave function $-x \log x$.) Shannon explained relation (8) as expressing the property of H that the uncertainty about one variable is never increased by knowledge of another. This, according to him, was among the properties that gave H its real justification. The idea is surely very appealing and pleasing, but it is not correct.

There are situations where knowledge about one variable in fact increases uncertainty about another. An example is given by Aczél and Daróczy (1975): assume that the probability of a raven being black versus white is 0.99 versus 0.01, but that if a raven has a white mother, the probability of it being white is 0.5. Then if we learn that a particular unobserved raven has a white mother the uncertainty about its colour is increased. In fact information about a variable x may increase uncertainty about x itself: the entropy of the probability distribution for the location of my housekeys increases when I discover that they actually are not, as I held to be very probable, in the pocket of my coat (cf. Uffink, 1990).

Seidenfeld (1979) discussed similar examples of 'reverse ordering" as important objections against the MEP. They show that entropy and information do not always vary in the same sense and that Shannon's explanation of relation (8) was not correct. This relation says only that the entropy of x is not *expected* to increase upon knowledge of y.

In view of the present puzzles Jaynes wrote:

> "... one can easily invent situations where acquisition of a new piece of knowledge (that an event previously considered improbable had in fact occurred) can cause an increase in the entropy."

and he concluded:

> "This paradox shows that 'information' is an unfortunate choice of word to describe entropy expressions." (Jaynes 1957b, p. 186)

This conclusion is rather disappointing, when compared with the earlier quotation in which it was argued that every measure of information except H will lead to contradictions. If now H itself leads to paradox when it is used as an information measure, one may very well ask again whether alternative measures would behave so much worse. We shall take up the question again after a discussion of the problems encountered with the extension to continuous probability distributions.

In conclusion, the strategy of justifying the MEP by means of Shannon's uniqueness theorem seems to fail. First of all, the assumptions on which the theorem rests are not all compelling, but contain also conventional elements. Secondly there is still the problem of extension to the continuum. Thirdly there are examples showing that the entropy expression has properties which do not correspond completely with what one would intuitively expect of an information measure. To be sure, these examples are perhaps more telling against intuition than against the Shannon entropy. But they do undermine the claim that the assumptions needed to characterize this expression uniquely can be justified by an appeal to consistency or intuitive assent.

5 Extension to the continuum: maximum relative entropy

Up till now we have assumed that the set S of possible cases is finite. If we wish to extend the MEP to more general situations, the first obvious case to consider is that where S is infinite but still discrete, e.g. $S = \mathbb{N}$. This case poses the problem that the entropy expression

$$H = -\sum_{i=1}^{\infty} p_i \log p_i$$

is now unbounded from above. It may occur that H becomes infinite for more than one distribution allowed by the constraints and thus does not possess a unique maximum. The problem how the MEP is to be applied in this case is not often studied in the literature, presumably because it seems almost neglible compared with those encountered in the extension to the continuum. We shall therefore leave it as it is, and simply assume that the constraints are such that a unique maximum for the entropy is admitted.

A more formidable problem is obtained when we assume that the set of possible states x forms a continuum, say $S = \mathbb{R}$. It is well known that the entropy expression (1) does not have a natural extension to this case. Indeed, the expression one would naively write down for this case,

$$-\int p(x) \log p(x)\, dx$$

for a probability density $p(x)$, has properties which are rather different from those of its discrete counterpart (1). In particular, probability densities mostly carry a physical dimension (say probability per length) which gives H the unit of "log cm", which seems somewhat odd. Also, in contrast to (1), this expression is not invariant under a reparametrization of S, e.g. by a change of unit. Further, H may now become negative, and is not bounded from above nor below so that new problems of definition appear. (Cf. Hardy, Littlewood and Pólya, 1934, p. 126.)

These problems are clarified if one considers how to construct an entropy for a continuous probability distribution starting from the discrete case. A natural approach is to

partition S into a disjoint subsets $(A_1, \ldots, A_n)$, and then calculate the entropy of the discrete probability distribution $P(A_1), \ldots P(A_n)$. The entropy of the continuous probability distribution ought then to be obtained by taking the limit of finer and finer partitions. Unfortunately, this approach is frustrated, because this limit is infinite for all continuous probability distributions. This divergence is also obtained –and explained– if one adopts the well-known interpretation[4] of the Shannon entropy as the least expected number of yes/no questions needed to identify the value of x, since in general it takes an infinite number of such questions to identify a point in continuum. In view of these problems many authors have denied the possibility of defining entropy expressions for the continuum.

A more fruitful way of dealing with the continuum is by replacing the entropy expression (1) by the so-called *relative* entropy. (See Kolmogorov, 1957; Gelfand, Yaglom and Kolmogorov, 1958; and Kullback 1957). For the discrete case, this entropy is defined as

$$H_{disc}(p, \mu) = -\sum_i p(x_i) \log \frac{p(x_i)}{\mu(x_i)} \tag{9}$$

where $\mu(x_i)$ are positive weights determined by some 'background measure' μ. In the special case where μ is the counting measure, i.e. if $\forall i : \mu(x_i) = 1$, the relative entropy (9) becomes equal to the (absolute) entropy (1). This relative entropy (not to be confused with the conditional entropy (7)!) however has a natural extension to the continuous case. The important difference with the absolute entropy (1) is that if one now partitions the real line in increasingly finer subsets, the probabilities $P(A_i)$ and the background weights $\mu(A_i)$ are both split simultaneously and the logarithm of their ratio will generally not diverge. In fact, it can be shown that this relative entropy is non-increasing under refinement:

$$-\sum_i P(A_i) \log \frac{P(A_i)}{\mu(A_i)} \;\geq\; -\sum_j P(B_j) \log \frac{P(B_j)}{\mu(B_j)}$$

if the partition $(B_1 \ldots B_m)$ is a refinement of $(A_1 \ldots A_n)$. Hence the relative entropy over a continuum can be defined unambiguously as the limit under increasing refinement if we make the assumption that μ does not vanish on any set A for which $P(A) > 0$.[5]

The relative entropy of the continuous probability measure P with respect to a background measure μ can be written as

$$H(P, \mu) = -\int \frac{\partial P}{\partial \mu}(x) \log \frac{\partial P}{\partial \mu}(x)\, d\mu(x) \tag{10}$$

where $\partial P / \partial \mu$ denotes the Radon-Nykodim derivative. In the case where μ is the Lebesgue measure λ, this reduces to

$$H(P, \lambda) = -\int p(x) \log p(x)\, dx \tag{11}$$

where $p(x)$ is the probability density of the probability measure P. This is of course exactly the expression one would have expected as an analogue of (1) also from a naive point of view. It is important to note, however, that because the relative character, i.e. the dependence on a second measure is brought out in (11), the expression is now invariant

[4] This interpretation supposes that the logarithm in (1) has base 2.

[5] Of course the limit may still be infinite (positive or negative) for particular choices of P and μ.

under a reparametrization of the continuum, as long of course as the background measure is unchanged.

In the case where μ is itself a probability measure, it is sometimes more convenient to represent both measures by their corresponding densities. I.e. if we put $p = \partial P/\partial\lambda$, $m = \partial\mu/\partial\lambda$, we get

$$\frac{\partial P}{\partial \mu}(x) = \frac{p(x)}{m(x)} \quad \text{and} \quad d\mu(x) = m(x)dx,$$

so that the expression (10) takes the form

$$H(p, m) = -\int p(x) \log \frac{p(x)}{m(x)}\, dx \tag{12}$$

which we shall use later.

It appears that in order to define an entropy for continuous probability distributions we should replace the concept of absolute entropy by that of relative entropy. How does this affect the maximum entropy principle? Obviously this can now be generalized into the *maximum relative entropy principle* (MREP): choose that probability distribution which, under given constraints, and for a given background measure μ maximizes the relative entropy. Technically, this fixes the problem of dealing with the continuum. The relative entropy does not change when we reparametrize x, because P and μ will transform in exactly the same way. Thus the MREP does not fall prey to the Bertrand paradox.

The new rule, however, is obviously different from the earlier (absolute) maximum entropy principle because it is relative to a choice of the background measure. Different choices of μ will lead to different probability assignments. So how do we choose μ? This in turn depends on how one interprets this measure. There are two options in the literature. In Jaynes' approach (Jaynes, 1968, 1973), μ is taken to reflect the physical symmetries in the problem at hand. In particular, if there is a physical symmetry group for the problem, μ must be invariant under the action of this group. For example when x is a location in space, and if the problem is symmetrical under spatial translations, the background measure must be proportional to the Lebesgue measure. Note that, as this example shows, μ need not be a probability measure.[6]

A second interpretation of the background measure is that it too represents a probability distribution, a *prior* distribution that corresponds to our knowledge of the system before the information encapsulated in the constraints comes in. In this interpretation the principle of maximum relative entropy becomes a rule for changing or updating a previous probability distribution. This version of the principle is used by Williams, Shore and Johnson, Skyrms, Van Fraassen and many others. It is also called the minimum cross-entropy or minimum information principle. It seems that Jaynes always rejected this point of view.

In both versions, the transition from an absolute to a relative entropy principle has important implications also for the first objection discussed in section 2, the '*ex nihilo nihil*'. This is not to say that the MREP is acceptable for those who interpret probability as an objective quantity. But it seems to me that the misgivings that many 'candid persons' have against the PIR or the MEP are connected with the suspicion that it just plucks the probability values out of thin air (cf. Edwards, 1972). Since the MREP requires a specification of the background measure as an extra mathematical ingredient, the situation is different, as follows.

[6] Unless, following Jeffreys, one allows for unnormalized (improper) probability distributions.

In Jaynes' interpretation a connection is made with the symmetries of the problem. It is, of course, still true that these symmetries characterize our state of knowledge rather than the physical world. But Jaynes (1973) also assumes that every circumstance that may 'exert an influence' is explicitly included in the statement of the problem. Hence the symmetries in our formulation of a problem are assumed to be reflections of those existing in the physical world. This means that the choice of the background measure presupposes fallible, empirical knowledge, and one does not proceed *ex nihilo*.

In the second interpretation the change in status of the entropy principle is even greater. Here, maximization of relative entropy is no longer regarded as a general principle by which one *assigns* a probability distribution. On the contrary, it should be regarded as a rule to *adapt* or *update* a distribution already in our possession. It has been aptly called a rule of 'probability kinematics'. And just like ordinary kinematics, the rule has to supplemented by a specification of initial conditions in order to obtain tangible results. This means that the question of how to asign values to probabilities, with which we started our discussion, is left to be answered by other means. And with this more modest goal, the second version of the MREP is also not vulnerable to the objection that it plucks the values of probabilities out of thin air.

Either way, there seems to be genuine and important progress in the extension of the MEP to the MREP. But there still remains the question of justification: why maximize relative entropy, and not, say, the relative analogue of (6)

$$\tilde{H}(P,\mu) = \int \phi(\frac{\partial P}{\partial \mu})\, d\mu \tag{13}$$

for some other concave function ϕ? One approach to answering this question could be to find a set of assumptions which characterize the relative entropy uniquely, in analogy with Shannon's theorem. Such a characterization has been given by Hobson (1971), but the assumptions needed seem to have less intuitive appeal than those of Shannon. Another axiomatization has been given by Rényi (1962). An alternative axiomatization, characterizing all expressions (13) is given by Uffink (1990). We shall not pursue this approach here. In the following we shall see that there are desirable properties obeyed only by a special class of concave functions ϕ. The choice $\phi(x) = -x \log x$, leading to the relative entropy, is among them but it is not unique.

6 Justification by consistency: Shore and Johnson

The strongest and most careful attempt to answer the question why the rule to maximize the relative entropy (12) is to be preferred was made by Shore and Johnson (1981). These authors explicitly present their work as a proof that this rule represents the unique correct rule of inference, and as a vindication of Jaynes' claim that every other rule will lead to contradictions. We shall see, however, that their results actually provide considerable evidence against these claims.

In the present approach, it is assumed that one is looking for a procedure by which a prior probability density $p(x)$ is changed into a posterior density $q(x)$ when new information is taken into account. Thus, we are dealing here with the second interpretation of the MREP discussed in the previous section, as a rule of updating. It is assumed that the new information specifies a set $\mathcal{I}$ of probability densities in which the posterior is constrained

to lie. For example, the constraint might fix the expectation value of some function f:

$$\mathcal{I} = \{q \ : \ \int f(x)q(x)\,dx = c\}$$

or puts inequality bounds on such expectations:

$$\mathcal{I} = \{q \ : \ a \leq \int f(x)q(x)\,dx \leq c\}.$$

But more general constraints are also allowed, e.g. fixing a conditional expectation:

$$\mathcal{I} = \{q \ : \ \int f(x)q(x|S_i)\,dx = c\}$$

where $q(x|S_i)$ is the conditional density restricted to a subset $S_i \subset S$.

$$q(x|S_i) = \begin{cases} \frac{q(x)}{\int_{S_i} q(x)dx} & \text{if } x \in S_i \\ 0 & \text{otherwise} \end{cases}$$

It is further assumed that the procedure takes the form of maximizing some relative uncertainty expression of the form $F(q,p)$ under the constraint $q \in \mathcal{I}$. However, the procedure is characterized not by desiderata on F, but by how the posterior depends on the prior distribution and the new information.

This approach, which differs considerably from that of relying on Shannon's uniqueness theorem, indeed solves two of the drawbacks which we found in section 4. In the first place, Shore and Johnson axiomatize the inference rule itself, instead of the uncertainty measure. The merit of this is that the question what convention we shall choose to scale the measure of uncertainty no longer plays any role. The rules 'maximize $H(q,p)$' and 'minimize $\exp -H(q,p)$' can be identified, because they yield the same result on the same input. Secondly, Shore and Johnson characterize the MREP rather than the MEP. The latter appears only as the special case where the prior distribution is uniform. Thus, no problems with the extension to the continuum appear. Both aspects represent important advantages. A weak point, of course, is still that one simply *assumes* in this approach that the inference rule proceeds by maximizing some functional depending only on prior on posterior.

Technically, the problem is formulated as follows. A system has a set S of possible states with an unknown true[7] probability density $q^\dagger$. We write the class of all probability densities over S as $\mathcal{P}$. The prior distribution, $p(x)$ represents a (subjective) estimate of $q^\dagger$ before new information is given. It is assumed that the prior is diffuse, $\forall x \in S : p(x) > 0$. The new information I is assumed to single out a closed[8] convex subset $\mathcal{I}$ of $\mathcal{P}$ in which the true probability $q^\dagger$ must fall. This is written as $I = (q^\dagger \in \mathcal{I})$. In response to this information, the prior density p is changed into a posterior density q in $\mathcal{I}$. It is assumed that the inference rule yields this posterior q as a function depending only on the prior p and the constraint I, symbolically written as:

$$q = I \circ p \tag{14}$$

[7]The assumption of the existence of a 'true' probability distribution may not be palatable to strict followers of the subjectivist view of probability. One can however easily replace the idea of a 'constraint on $q^\dagger$' by a 'constraint on the posterior q' without damage to the mathematical argument.

[8]In Shore and Johnson (1981) this closure was specified as being understood in L^1-norm.

where $\circ$ is an 'updating operator'. Shore and Johnson give five "consistency axioms" for this updating operator (Shore and Johnson, 1980; Johnson and Shore, 1983).

"1. Uniqueness: The result should be unique."

This axiom is in fact already implicit in the notation (14), specifying that q is determined as a function of I and p.

"2. Invariance: The choice of coordinate system should not matter."

Thus, if $\Gamma : p(x) \rightarrow \hat{p}(y) = p(x(y))|\frac{dx}{dy}|$ represents the transformation of a probability density under a bijective reparametrization of the set S, one has

$$\Gamma(I \circ p) = (\Gamma I) \circ (\Gamma p)$$

where ΓI is to be read as stating that the true distribution $\Gamma q^{\dagger}(y)$ obeys the transformed constraint $\Gamma \mathcal{I} = \{q \in \mathcal{P} \ : \ \Gamma^{-1} q \in \mathcal{I}\}$. This expresses the desire to avoid Bertrand's paradox.

"3. System independence: It should not matter whether one accounts for independent information about independent systems separately in terms of different densities or in terms of a joint density."

Thus, let p_i and I_i denote prior probability densities and constraints for two systems, each with its individual state x_i, $i = 1, 2$. For independent systems, a probability density describing the combined system should take the form of a product of the separate densities. Thus, when the prior joint density is taken as $p(x_1, px_2) = p_1(x_1)p_2(x_2)$, the axiom states:

$$(I_1 \wedge I_2) \circ (p_1 p_2) = (I_1 \circ p_1)(I_2 \circ p_2)$$

where $I_1 \wedge I_2$ is the conjunction of I_1 and I_2.

"4. Subset independence: It should not matter whether one treats disjoint[9] subsets of system states in terms of separate conditional densities or in terms of the full density."

This requirement may need some explanation. Note that the axiom differs from the preceeding one, at least in spirit, in the sense that one does not consider the structure of a composite system but the structure of the state space. The motivation behind the axiom can perhaps be illustrated as follows. Suppose one is interested in the probability of different political parties winning the next election. The set of parties S is divided into the subsets S_L and S_R (left and right wing). Now suppose some partial information I_L is obtained (say by a poll) that indicates that among left-wing voters a relative shift in support is to be expected, say from radical left to social-democrat. Let the updated probability distribution of party x winning the election be written as $q(x) = I_L \circ p(x)$. Now it seems reasonable to argue that since the information I_L concerns parties from the left only, the updating procedure should affect the distribution over that part of the political spectrum only.

[9]The original text reads 'independent' instead of 'disjoint'. It seems clear from the context however that the latter term was intended.

Thus if we conditionalize the updated probability $q(x)$ under the supposition that the winning party belongs to the set S_L of left-wing parties, this should give the same result as when we update the prior distribution conditionalized on S_L:

$$q(x|S_L) = I_L \circ p(x|S_L).$$

Similarly, conditionalizing the updated probability under the supposition that the winner belongs to S_R should equal the updated conditional probability $p(x|S_R)$. Furthermore, this distribution should not be affected by the information I_L concerning left-wing parties only and remain equal to the prior:

$$q(x|S_R) = I_L \circ p(x|S_R) = p(x|S_R)$$

More generally, let $S_1, \ldots S_m$ denote disjoint subsets of S and let each of the informations I_j merely constrain the conditional distributions $q(.|S_j)$. Then, first updating the prior distribution under $I_1 \wedge \ldots \wedge I_m$ and next conditionalizing on S_j should lead to the same result as first conditionalizing the prior density and next updating under the constraint I_j. Formally:[10]

$$\text{If } \; q = (I_1 \wedge \ldots \wedge I_m \circ p) \; \text{ then } \; q(x|S_j) = I_j \circ p(x|S_j) \tag{15}$$

The final axiom is:

5. "In the absence of new information, we should not change the prior."

Thus, when we are given the trivial constraint $I = (q^\dagger \in \mathcal{P})$ then $I \circ p = p$.

The above axioms do indeed seem reasonable for the outlined problem. This is not to say, however, that they are compelling or motivated by consistency. The first demand, for example, rather seems to follow from the desire that the rule *settles* the problem. A rule that in some occasions leaves more than one option open would only be less useful; i.e. it would have to be supplemented by other considerations. (To give an example from a quite unrelated area, the traffic regulation laws do not prescribe uniquely how one has to drive. But that does not mean these laws are inconsistent.) Actually, the MREP rule would not meet axiom 1 in general either, if one dropped the earlier assumption that the constraints pick out a closed convex set of probability distributions.[11]

Still, the Shore-Johnson axioms are to a large extent reasonable. However they do not characterize the MREP uniquely. The following theorem is shown in the appendix of this paper.

Theorem 2 *An updating procedure satisfies the five consistency axioms above if and only if it is equivalent to the one of the rules*

$$\textit{Maximize } \; U_r(q,p) \quad \textit{under the constraint } I. \tag{16}$$

[10] This formulation is what Shore and Johnson (1981) call *weak* subset independence. A stronger version than this is actually necessary for the proof of their theorem, namely the demand that $p(.|S_j)$ is also unaffected by any information M which merely specifies the overall probabilities of the sets S_j. Thus, if $M : (q^\dagger(S_j) = \alpha_j, j = 1, \ldots m)$ then $(I_1 \wedge \ldots \wedge I_m \wedge M) \circ p(.|S_j) = I_j \circ p(.|S_j)$.

[11] To see what goes wrong with the MREP if we allow non-convex sets it is sufficient to think of the case where the constraint fixes the value of the relative entropy, i.e.: $I = \{q : H(q,p) = c\}$. An example showing that there may be no solution when the constraint set is convex but not closed is given by Csiszár (1985): let p be a standardized normal (Gaussian) probability density and $I = \{q : |\langle x^3 \rangle| \leq 1\}$. Here $H(q,p)$ can be made arbitrarily close to zero, without attaining this value for any $q \in I$. See also the discussion by Williams (1980).

where

$$U_r(q,p) = \left(\int (\frac{q(x)}{p(x)})^r q(x)\, dx \right)^{-1/r} \tag{17}$$

and $r > -1$.

The MREP rule is a member of this class of procedures if we define

$$U_0(q,p) := \lim_{r \to 0} U_r(q,p)$$

because a Taylor expansion in r gives (cf. Hardy, Littlewood and Pólya, 1934, p. 15):

$$\begin{aligned} U_r(q,p) &= \exp\left(\frac{-1}{r} \log \int (\frac{q}{p})^r q\, dx \right) \\ &= \exp\left(\frac{-1}{r} \log \left(1 + r \int q \log \frac{q}{p}\, dx + \mathcal{O}(r^2) \right) \right) \\ &\to \exp H(q,p) \text{ when } r \to 0 \end{aligned}$$

For fixed densities q and p, U_r is a non-increasing left-continuous function of r, provided one defines the limiting cases as [12]

$$\begin{aligned} U_\infty(q,p) &= \left(\sup_x \frac{q(x)}{p(x)} \right)^{-1} \\ U_{-1}(q,p) &= P(\{x \ : \ q(x) > 0\}) \\ U_{-\infty} &= \left(\inf_x \frac{q(x)}{p(x)} \right)^{-1} \end{aligned}$$

For general $r > -1$, the expressions $\log U_r$ are called Rényi entropies (Rényi, 1962). The case $r = -0.5$ is remarkable as the only one in which $U_r(q,p)$ is symmetrical in p and q, and which determines a distance function between probability distributions (See Bhattacharrya, 1943; Wootters, 1981; Hilgevoord and Uffink 1991).

Clearly the axioms given above do not entail the conclusion of Shore and Johnson that the MREP is the uniquely consistent rule of inference. Something is still missing in the argument. What is it? The answer to this question appears when the informal statement of the axioms is compared with the exact formulation which is used in the proof. In particular, when the system independence axiom is formulated exactly, something much stronger than the informal statement is demanded. Let us spell this out.

Suppose there are two systems, 1 and 2, with possible states $x_1 \in S_1$ and $x_2 \in S_2$ and two prior distributions $p_1(x_1)$ and $p_2(x_2)$. The unknown true probability densities are $q_1^\dagger$ and $q_2^\dagger$ respectively, and there are two separate pieces of information

$$I_1 = (q_1^\dagger \in \mathcal{I}_1) \text{ and } I_2 = (q_2^\dagger \in \mathcal{I}_2)$$

[12] Note that if we replace the prior probability measure by the Lebesgue measure, as in eq. (13), the expressions U_∞, U_0 and U_{-1} become familiar measures of width of the density $q(x)$, U_∞ being the 'equivalence width' and U_{-1} the support. The supremum in U_∞ is to be understood as the 'essential supremum', i.e. as the least upper bound of the values s such that for all $\epsilon > 0$ the set $A_\epsilon = \{x \ : \ q(x)/p(x) > s - \epsilon\}$ has positive prior measure: $P(A_\epsilon) > 0$. The essential infimum is defined similarly.

where the constraint sets $\mathcal{I}_1, \mathcal{I}_2$ are closed convex subsets of $\mathcal{P}_1$ and $\mathcal{P}_2$ respectively. The information I_1 and I_2 can be processed either for each system separately to obtain the posteriors

$$q_1 = I_1 \circ p_1 \quad , \quad q_2 = I_2 \circ p_2 \tag{18}$$

or, alternatively, we can think of the conjunction of I_1 and I_2 as one piece of information pertaining to a combined system composed of 1 and 2 and described by a joint probability density.

It is now assumed first that the systems are in fact independent, i.e. the true joint density is given by the product

$$q^\dagger(x_1, x_2) = q_1^\dagger(x_1) q_2^\dagger(x_2) \tag{19}$$

The joint information about the combined system is then formulated by the constraint $I_1 \wedge I_2 = (q^\dagger \in \mathcal{I}_{12})$ where:

$$\mathcal{I}_{12} = \{q(x_1, x_2) \ : \ \int q(x_1, x_2)\, dx_2 \in \mathcal{I}_1 \text{ and } \int q(x_1, x_2)\, dx_1 \in \mathcal{I}_2\}$$

is the set of all joint distributions q of which the two marginals fall in $\mathcal{I}_1$ and $\mathcal{I}_2$ respectively. Using the rule of inference on the joint prior $p_1(x_1)p_2(x_2)$ under this constraint leads to a posterior $q(x_1, x_2)$ represented by

$$q = (I_1 \wedge I_2) \circ (p_1 p_2)$$

which we can compare with the product $q_1 q_2 = (I_1 \circ p_1)(I_2 \circ p_2)$ of the posteriors which result from (18). At this point Shore and Johnson state:

> "Because p_1 and p_2 are independent and because I_1 and I_2 give no information about any interaction between the two systems, we expect these two ways to be related by $q = q_1 q_2$, *whether or not* [(19)] *holds.*" (p. 29, emphasis added)

It is this emphasized phrase that goes much further than the original statement of axiom 4 and that excludes the generalized rules (17) for $r \neq 0$. The phrase means that the axiom of system independence is intended to hold not only for independent systems but regardless of whether the systems are independent! In their article of 1981, Shore and Johnson made the point even more explicit:

> "Whether the systems are in fact independent is irrelevant; the property [of system independence] applies as long as there are independent priors and independent new information." (p. 475)

Clearly we should raise the question whether this remarkable addition can be explained as a consistency requirement. Let us first consider the reasons Shore and Johnson themselves offer in the above quotations. Can we use the fact that the joint prior is the product of independent priors p_1 and p_2 as a motivation? I don't think so. After all the prior density is in this approach just an estimate of $q^\dagger$ before the new information comes in. Thus, it need not be based on extensive knowledge of the two systems. One might very well choose a factorizing prior $p(x_1, x_2) = p_1(x_1)p_2(x_2)$ even when one knows the two systems to be correlated, but does not know how they are correlated. Certainly the factorization of the prior is no ground for the demand that the posterior should also factorize. So let

us consider the second argument, that I_1 and I_2 give no information about any interaction or correlations between the two systems. This, of course, refers to the fact that I_1 and I_2 pertain to the systems 1 and 2 separately, and thus by themselves do not convey information about interaction or correlations between the systems. But no information about interaction is not the same as the information that there is no interaction! So this is not sufficient either for the demand that after the reception of I_1 and I_2 we should still regard the systems as independent.

Let me give two examples to show that the extra requirement is not only unwarranted but can actually be unreasonable. Suppose the two systems are two inhabitants of Utrecht. They can be in either of two states, 'brown-eyed' or blue-eyed'. Knowing nothing in particular about these persons and judging that blue eyes and brown eyes are about equally prolific, I would opt for a prior that gives equal probability $\frac{1}{4}$ to all four combinations. Now suppose information I_1 specifies the exact time and place of birth of the first person.[13] This, certainly, gives no information about the second person, or about any correlation between them. So, I update the marginal distribution for system 1. Similarly, let I_2 specify the exact time and place of birth of system 2. In fact, not expecting any correlation between eye colour and birth date, I keep the two posteriors exactly the same as the priors. But now suppose that it turns out that the two persons were born the same day, and in the same place. Then I could reason very differently when the information is combined: it would raise my suspicion that they might be twins, so that I expect some correlation between the colours of their eyes. Such an expectation would surely not be 'inconsistent'.

The second example is in the same spirit, and is offered mainly to show that the objection can be put in exact mathematical form. The mathematics is borrowed from the famous Einstein-Podolski-Rosen argument in quantum mechanics. I wish to emphasize, however, that the issues of non-locality usually associated with this argument are irrelevant for our purpose. Also, our discussion will focus on the parameters of the macroscopical apparatus and not on the microscopic particles.

In an EPR-experiment, two photons are prepared in a singlet state and then fly apart towards two detectors *Left* and *Right*. In front of these detectors there are two coplanar polarization filters oriented in direction θ_1 and θ_2 ($0 \leq \theta_i < 2\pi$). Suppose these directions are initially regarded as completely unknown, so that the prior distribution over θ_1, θ_2 is given by

$$p(\theta_1, \theta_2) = \frac{1}{(2\pi)^2}.$$

When the photons reach the filters they are either absorbed or transmitted and registered by the detector. So for each detector L and R the outcome will either be '+' (detection) or '−' (absorption by the filter), and one of the four combinations $(+,+)$, $(+,-)$, $(-,+)$, or $(-,-)$ is obtained experimentally. The probability distribution for these joint outcomes depends on θ_1 and θ_2 and is specified by quantum theory as:

$$\begin{pmatrix} p_{\theta_1\theta_2}(++) & p_{\theta_1\theta_2}(+-) \\ p_{\theta_1\theta_2}(-+) & p_{\theta_1\theta_2}(--) \end{pmatrix} = \begin{pmatrix} \frac{1}{2}\cos^2(\theta_2-\theta_1) & \frac{1}{2}\sin^2(\theta_2-\theta_1) \\ \frac{1}{2}\sin^2(\theta_2-\theta_1) & \frac{1}{2}\cos^2(\theta_2-\theta_1) \end{pmatrix} \tag{20}$$

The question is how to change the prior distribution for the unknown orientations in the

[13] I admit that in this example it is not easy to say how this information singles out a convex subset of probability distributions which is closed in L^1 topology.

cases when the data consist of

$$\begin{aligned} D_1 &: \quad + \text{ is obtained in detector } L \\ D_2 &: \quad + \text{ is obtained in detector } R \\ D_1 \wedge D_2 &: \quad (+,+) \text{ is obtained in both } L \text{ and } R \end{aligned}$$

Let us first see what a traditional method of inference would give. In Bayesian statistics one would update the prior distribution as

$$q_{D_j}(\theta_1, \theta_2) = p(\theta_1, \theta_2 | D_j) \propto p(D_j | \theta_1, \theta_2) p(\theta_1, \theta_2)$$

for $D_j = D_1, D_2$ or $D_{12} = D_1 \wedge D_2$ respectively. The result is

$$p(\theta_1, \theta_2 | D_1) = p(\theta_1, \theta_2 | D_2) = \frac{1}{(2\pi)^2}$$

i.e. neither of the two results '+ in L' or '+ in R' by themselves give any reason to change our opinion. However,

$$q(\theta_1, \theta_2 | D_1 \wedge D_2) = \frac{1}{2\pi^2} \cos^2(\theta_1 - \theta_2)$$

so that the combined data give some reason to believe that the directions are parallel, and strong evidence against believing that they are orthogonal.

The challenge for this example is, of course, to show that data as considered here can be represented as convex sets of probability distributions, as assumed in the discussion of Shore and Johnson. A simple way to achieve this is by letting each datum D_j correspond to the set of all marginal posteriors $q(.|D_j)$, which can be obtained by Bayesian conditionalization from arbitrary priors. Thus:

$$\mathcal{I}_1 = \{q(\theta_1) \ : \ \exists \rho(\theta_1, \theta_2) \ \text{ such that } \ q(\theta_1) \propto \sum_b \int p(+, b | \theta_1, \theta_2) \rho(\theta_1, \theta_2) \, d\theta_2\},$$

$$\mathcal{I}_2 = \{q(\theta_2) \ : \ \exists \rho(\theta_1, \theta_2) \ \text{ such that } \ q(\theta_2) \propto \sum_a \int p(a, + | \theta_1, \theta_2) \rho(\theta_1, \theta_2) \, d\theta_1\}$$

$$\mathcal{I}_{12} = \{q(\theta_1, \theta_2) \ : \ \exists \rho(\theta_1, \theta_2) \ \text{ such that } \ q(\theta_1, \theta_2) \propto \int p(+, + | \theta_1, \theta_2) \rho(\theta_1, \theta_2) \, d\theta_1\}$$

where $a, b \in \{+, -\}$. In the assumed theoretical model (20) the sets $\mathcal{I}_1, \mathcal{I}_2$ are in fact equal to the entire sets $\mathcal{P}_j$, corresponding to the fact that whatever our prior belief about θ_j was, the result found in a single detector will not change it.

The examples show, I believe, that evidence for a dependence or correlation between two systems can very well be contained in the mere logical conjunction of two pieces of information, each of which taken separately give no clue for the dependence at all. This means that the additional requirement of Shore and Johnson, namely that the 'system independence axiom' should hold regardless of whether the systems are independent, is unreasonable. However, without this addition, the MREP is not the unique rule of inference that satisfies their consistency axioms.

Finally some remarks on how related inference rules violate the Shore-Johnson axioms. First, consider the rule to maximize (17) with $r \leq -1$. This case differs essentially from

that of $r > -1$ because now x^{1+r} is convex and $x^{-1/r}$ is increasing, so that the rule is not of the form of maximizing an expression (6) for the concave function $\phi(x) = x^{1+r}$ (as for $-1 < r < 0$) or a minimization with ϕ convex, as for $r > 0$. Thus, for $r < -1$, (17) has radically different properties. However the rules to *minimize* U_r when $r \leq -1$ do satisfy all the axioms except the first: if I demands that $q(x)$ equals zero within a region of positive prior probability, then $U_r(q,p) = 0$ if $r < -1$ for all $q \in \mathcal{I}$. Then there will not be a unique solution to the minimization problem. (A similar problem in the case $r > -1$ may occur if the density $q(x)$ is constrained to have an integrable singularity.) Next, the rules to maximize the more general expressions (13) for concave ϕ violate both system independence and the strong subset independence axioms (cf. footnote 10). However, these rules do obey the weak subset dependence axiom 4 as expressed by equation (15). It seems likely that for strictly concave ϕ they are the only rules obeying axioms 1,2, 5 and weak 4. A more detailed study is given in Uffink (1990).

7 Justification by consistency: Tikochinsky, Tishby and Levine

Another claim that the principle of maximum entropy can be proven to be the only method of statistical inference satisfying conditions of consistency was made in 1984 by Tikochinsky, Tishby and Levine (TTL) (1984a, 1984b). In the approach of these authors, like that of the previous section, conditions are imposed only on the inference scheme itself and not on the entropy expression. The approach differs from Shore and Johnson in the sense that the discussion is restricted to discrete probability distributions and the original maximum (absolute) entropy principle instead of the MREP. That is, we are dealing here with a rule for assigning probability values and not with one for updating a prior probability assignment. Also, only linear constraints are considered. Important advantages are, however, that instead of the five axioms of Shore and Johnson only two very simple conditions are imposed here which were baptized 'repetition consistency' and 'uniformity'. Furthermore, TTL dispense with the implicit assumption of Shore and Johnson that the inference procedure is obtained by maximizing some functional of the probability distribution.

The argument or TTL was hailed by Skilling as "brilliantly simple" (Skilling 1984a) and "deeply compelling" (Skilling 1984b) and he concluded that:

> "These ideas justify the fundamental claims made for maximum entropy in data analysis. It is sufficient to know that we must use maximum entropy – or lay ourselves open to the charge of inconsistency. Let's get on with it." (Skilling 1984a)

However, critical comments were published by Johnson and Shore (1984) and by Shimony (1984). In particular, Johnson and Shore agreed that the condition of repetition consistency was "extremely compelling", but they pointed out that the condition of uniformity was not formulated precisely and that a crucial step in the proof remained unclear. Nevertheless, they concluded that the theorem was probably correct, even if the proof appeared to be flawed. Shimony likewise did not doubt the validity of the theorem but argued that one of their assumptions (equation (22) below) need not hold universally. We shall show, however, that the theorem of TTL is false, even if their later clarifications (1984b) and replies (1984c, 1984d) are taken into account.

The problem is described by TTL as follows. Consider an experiment described by a probability distribution over a set of n possible outcomes $S = \{x_1, \ldots, x_n\}$. It is assumed that there is some algorithm which picks out a unique probability distribution for the experiment whenever the constraints

$$\langle f_k \rangle := \sum_{i=1}^{n} p_i f_k(x_i) = \alpha_k \tag{21}$$

are imposed, where $k = 1, \ldots m$, $m \leq n-1$. It is further assumed that the experiment can be repeated under identical conditions. We can then argue in two ways.

(i). First we apply the algorithm with the constraint (21) to obtain a unique distribution $p = (p_1, \ldots p_n)$ over S. Then we consider N independent (identically distributed) repetitions of the experiment under this probability distribution. Let $\vec{x} = (x_{i(1)}, \ldots, x_{i(N)}) \in S^N$ stand for the sequence of outcomes in this repeated performance. The probability of such a sequence is given by the product

$$P(\vec{x}) = p_1^{N_1} \cdots p_n^{N_n} \tag{22}$$

where N_i is the frequency with which outcome x_i occurs in the sequence $\vec{x}$ and $\sum N_i = N$. Now collect all sequences that differ only by a permutation, i.e. that show the same frequencies $\tilde{N} = (N_1, \ldots, N_n)$, to obtain a more condensed description. This gives the multinomial distribution

$$P(\tilde{N}) = \frac{N!}{N_1! \cdots N_n!} p_1^{N_1} \cdots p_n^{N_n} \tag{23}$$

for the probability of obtaining the frequencies $\tilde{N}$. Further, consider the average of the functions f_k, taken over the sequence of outcomes

$$\bar{f}_k(\vec{x}) = \frac{1}{N} \sum_{i=1}^{n} N_i f_k(x_i)$$

and notice that these averages depend on the observed frequencies only, and can thus be considered as functions of $\tilde{N}$:

$$\bar{f}_k(\vec{x}) = \bar{f}_k(\tilde{N})$$

The expectation values of $\bar{f}_k$ are, of course, equal to those of f_k, and thus we have

$$\langle \bar{f}_k \rangle := \sum_{\tilde{N}} \bar{f}_k(\tilde{N}) P(\tilde{N}) = \alpha_k. \tag{24}$$

(ii). We directly consider an N-fold independent repetition of the experiment, and regard it as a single compound experiment with the frequencies $\tilde{N}$ as possible outcomes. We impose the constraints (24) on the expectations of $\bar{f}_k(\tilde{N})$ and apply the algorithm to obtain a unique probability distribution $P'(\tilde{N})$.

Repetition consistency is now the demand that the results of procedures (i) and (ii) agree, i.e. we should have

$$P(\tilde{N}) = P'(\tilde{N}) \tag{25}$$

TTL claim to prove that the only algorithm obeying this consistency condition and which is 'uniform' is the maximum entropy principle. Here, the condition of uniformity is formulated as the demand that the algorithm treats "all data of the form [(21)] using

one and the same procedure." In particular it is urged that the value of n does not have a special standing in the algorithm. For example, a uniform algorithm may not use a different procedure depending, say, on whether n were prime or not (Tikochinsky, Tishby and Levine, 1984b and 1984c).

To see that the theorem is false, consider the absolute counterparts of the inference rules (17):

$$\text{"Maximize } \; U_r(p)\text{"} \tag{26}$$

where

$$U_r(p) = \left(\sum_{i=1}^{n} p(x_i)^{1+r} \right)^{-1/r} \tag{27}$$

and $r > -1$. Here, the only role of n is to provide the upper limit of the summation. Hence these rules are clearly uniform in the intended sense of the word, or at least no less so than the MEP itself. Yet they also obey repetition consistency because, for any distribution of the form (22),

$$U_r(P)^{-r} = \sum_{\vec{x} \in S^N} P(\vec{x})^{1+r} = \sum_{i(1)\ldots i(N)} \Big(p(x_{i(1)}) \cdots p(x_{i(N)}) \Big)^{1+r} = \left(\sum_i (p_i)^{1+r} \right)^N = U_r(p)^{-Nr}$$

so that maximization (or minimization) of $U_r(P)$ for the compound experiment is exactly equivalent to the maximization of $U_r(p)$ in a single experiment.

The error in the proof can be seen more clearly as follows. The space of all probability distributions for the compound experiment is typically of very high dimension. Indeed, as TTL point out, there are $l = \binom{N+n-1}{n-1}$ different sets of frequencies $\tilde{N}$, and thus the space of probability distributions over $\tilde{N}$ is $l-1$-dimensional (or l-dimensional if one treats the normalization condition as an extra constraint). TTL use their condition of uniformity to argue that if we impose the constraint (24) on this probability space the algorithm should yield a distribution in which n does not play a special role, "since n is not an input and is unknown to the problem in the l-dimensional space" (1984b). An important point however is that, by assumption, we are considering *independent* repetitions in both arguments (i) and (ii). Thus although the probability space for the compound experiment is indeed embedded in an l-dimensional probability space, we are actually only concerned with the subset corresponding to independent repetitions:

$$\{P(\tilde{N}) \; : \exists a_1, \ldots, a_n \;\; a_i \geq 0, \sum_i^n a_i = 1 \text{ such that } P(\tilde{N}) = \frac{N!}{N_1! \cdots N_n!} a_1^{N_1} \cdots a_n^{N_n}\},$$

i.e. a curved $n-1$-dimensional hypersurface in obvious one-to-one correspondence with the probability space of the original experiment. Thus the assumption of independence of the trials actually forces the parameter n to play a special role.

The condition of uniformity as applied by TTL ignores this and effectively demands that the algorithm, when applied to the constraint (24) on all distributions in the $l-1$-dimensional space, selects the same distribution as when it is applied to this constraint on the $n-1$-dimensional hypersurface of independent distributions. This means that in the case when it is *not* given that the repetitions of experiment are independent the algorithm should nevertheless reach the same result as when this *is* given! This is very similar to

the hidden requirement of Shore and Johnson, and we have already argued in the previous section that this is not reasonable.

It is interesting to note that the same set of inference rules allowed by the informal Shore-Johnson consistency axioms are also allowed by the TTL requirements. But it is hard to say whether they the only ones because, as noted by Shore and Johnson (1984), more tacit assumptions seem to be present in the present approach. From numerical examples one can see, however, that maximization of the more general expressions (6) need not obey repetition consistency.

8 The Judy Benjamin Problem

Van Fraassen (1981) proposed the problem of evaluating the merits of the MREP inference for the problem of updating a prior distribution over three possible disjoint events, A, B and C under a constraint of the form

$$\mathcal{I} = \{Q \ : \ Q(A|A \cup B) = \alpha\}.$$

He named this the 'Judy Benjamin problem', after a movie character. He argued that the solution yielded by the MREP needed more justification. Next, Van Fraassen, Hughes and Harman (1986) formulated a list of desiderata for general inference rules in this problem. They emphasized that these desiderata were not intended as compelling consistency requirements. They showed that the desiderata were satisfied by the MREP as well as two other rules which they called 'MTP' (maximum transition probability) and 'MUD'. They also argued that none of these rules was clearly superior to the others.

The desiderata read as follows:

> "1. If $\alpha = 1$ the prior is transformed by Simple [i.e. Bayesian] Conditionalization on $A \cup C$; if $\alpha = 0$ by Simple Conditionalization on $B \cup C$.
>
> 2. If α equals the prior conditional probability $P(A|A \cup B)$ then all probabilities stay the same.
>
> 3. The ratio $P(C)/P(A)$ should change (to $Q(C)/Q(A)$) by a factor $\gamma(s, r)$ which is a function only of the initial odds $s = P(B)/P(A)$ and the constrained odds $r = Q(B)/Q(A)$.
>
> 4. The function γ described in 3 is such that $\gamma(1/s, 1/r) = s/r\gamma(s, r)$ [This is relabeling invariance for the interchange of A and B.]" [14]

These principles are clearly related to the Shore-Johnson formulation. Thus, desideratum 2 is identical with axiom 5 of Shore and Johnson, and desideratum 3 is comparable in spirit to (though not identical with) their subset independence axiom. Only the system independence axiom, which would not be meaningful with only three possible cases, and the uniqueness axiom are left out here.

As one might already expect, and is easy to show, all the rules (17) fulfil the present desiderata. The special rules that Van Fraassen, Hughes and Harman call MTP and MUD

[14] I have changed the notation in this quotation. Van Fraassen, Hughes and Harman also formulate a fifth principle, which is not reproduced here because it is, as they make clear, actually already contained in the first desideratum.

correspond to the cases $r = -0.5$:

$$U_{-0.5}(Q, P) = \left(\sum_i \sqrt{p_i q_i} \right)^2 ,$$

and and $r = \infty$,

$$U_{\infty}(Q, P) = \max_i \frac{q_i}{p_i}.$$

I have not been able to determine whether the rules (17) are the only ones that obey the above desiderata. However this seems very likely to be the case. More general rules to maximize the expressions of the form (6) for arbitary concave ϕ can violate desideratum 3.

It is interesting to consider briefly the case of the rules to minimize U_r with $r < -1$. As noted before (section 6), $U_r(Q, P)$ then becomes zero as soon as the posterior is constrained to be zero on a set of non-zero prior probability. This happens in the present problem if α is 0 or 1. The rule to minimize U_r then becomes mute and does not yield a unique solution. Hence they do not reduce to Bayesian conditionalization and violate desideratum 1. But one can simply amend the rules by stipulating that simple conditionalization should take over whenever the minimum U_r distribution is not unique. Understood in this way, the case of $r \leq -1$ is allowed by the FHH conditions, in contrast to those of Shore and Johnson or TTL. (Of course this amendment will not help in more general constraints which can be considered when there are more then three possible events.)

As a conclusion to this section, it is gratifying to be able to give an explicit numerical solution to the original formulation of the Judy Benjamin problem (What should $Q(C)$ be when $P(A) = P(B) = \frac{1}{4}$, $P(C) = \frac{1}{2}$ and $\alpha = 0.25$?) by the result that she remains within the bounds set upon her by her creators if she chooses $Q(C)$ between 1/3 (using $r = -\infty$) and 0.6 ($r = \infty$).

9 Conclusions

The MEP evades the problems that beset the notorious Principle of insufficient reason. It is a concrete and clear recipe that yields unique numerical values for probabilities in many problems where more orthodox methods of inference do not. The most obvious problem facing this rich method is its justification, precisely because it yields unique probability values in cases where, according to orthodox methods, there is no unique solution. So why should one choose the maximum entropy probability distribution?

We have discussed three different approaches to solving this problem of justification by Jaynes, by Shore and Johnson, and by Tikochinsky et al. All these authors claim that the maximum entropy principle is justified as the unique consistent method of inference. We have found these approaches to be defective. Jaynes' approach puts a heavy weight on the assumptions of Shannon's uniqueness theorem, as if they were implied by the ideal of consistency itself. It has been argued that Shannon's assumptions simply cannot bear this weight. The approach of TTL fails because it rests on a technically false theorem. Shore and Johnson's work is the most sophisticated of the three approaches. However, in their analysis a hidden requirement is made, additional to their explicitly stated ones, which can be expressed as the demand that when it is not given whether systems (or experiments) are to be regarded as dependent we are justified in believing that they are independent. A similar conviction seems to lie at the bottom of the work of Tikochinsky et al. as well. This

sounds as a curious echo of the old principle of insufficient reason, but now on the level of dependence. I have argued by means of examples that the requirement is not reasonable.

It has been demonstrated that a viable class of inference rules do not obey the additional requirement. This class of 'Maximum (relative Rényi) Entropy' principles also satisfy a list of desiderata put forward by Van Fraassen et al. These rules can be seen as a new 'continuum of inductive methods', to use the terminology of Carnap, generalizing the maximum entropy method. It seems that more research would be needed to assess their performance in concrete cases and in general.

On the one hand one might ask whether there still are other properties of the Shannon entropy that justify its privileged status within this continuum. It is of interest that in his original article of 1957 Jaynes actually considered the expression $-U_1^{-1} = -\sum_i p_i^2$ as an alternative with "many of the qualitative propeties of Shannon's information measure, and in many cases leading to substantially the same results." However, he dismissed this expression for the reason that its maximum value under a linear constraint might be attained at the boundary of the constraint set, in which case it cannot be found by the Lagrange multiplyer technique.

One may indeed consider it desirable that an inference method should choose for a distribution from the interior of a constraint set, in order not to appear too biased. It is easy to show that this will hold for the Rényi entropies under linear constraints in case $r \leq 0$. Thus, this desideratum does not single out the Shannon entropy uniquely. Furthermore, one should note that once the generalization to arbitrary convex constraint sets $\mathcal{I}$ is accepted, all measures of uncertainty will occasionally attain their maximum at the boundary of the constraint set.

On the other hand, it should also be investigated how the present class of inference rules fares against even more general alternatives. This issue is particularly significant because some of the objections levelled against the Shannon entropy in section 4, namely that it does not always vary in the same sense as our information about a variable, holds equally for the more general entropy expressions.

It seems to early therefore to recommend the unqualified use of such rules. In fact another problem which affects all of them in equal measure is the question how to choose the constraint set. As mentioned in the introduction, this will be the subject of a sequel paper.

Acknowledgments

I thank Jasper Boessenkool, Tim Budden, Dennis Dieks, Fred Muller and Pieter Vermaas for helpful comments. It is a pleasure to thank Jeremy Butterfield and the other members of the Department of History and Philosophy of Science at Cambridge and Harvey Brown at the Subfaculty of Philosophy in Oxford for their hospitality and encouragements. This work was supported by a grant from the British Council and the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO).

10 Appendix

Here we show that an inference rule obeys the restricted Shore-Johnson axioms if and only if it is of the form (17). We start with the 'if' part.

It is obvious that U_r is invariant under any coordinate transformation, so that axiom 2 is evidently fulfilled. To show that axiom 1 is fulfilled, we show that U_r obeys the following property, called Schur convexity (see Roberts and Varberg, 1973; Uffink 1990):

$$\text{if} \quad U_r(q_1, p) = U_r(q_2, p) := U, \quad \text{then} \quad U_r(\alpha q_1 + \beta q_2, p) > U \tag{28}$$

for all $0 < \alpha < 1$, $\alpha + \beta = 1$. This means that whenever there are two distinct probability densities q_1 and q_2 allowed by the constraint set $\mathcal{I}$ with equal Rényi entropy, there will be a third density $\alpha q_1 + \beta q_2$ with higher Rényi entropy. Since the constraint set is assumed to be convex, this third density is also in $\mathcal{I}$. Hence, if there is a maximum, it is necessarily unique. To prove the inequality (28) we observe that for $r > 0, x \geq 0$, $f(x) = x^{1+r}$ is strictly convex and $g(x) = x^{-1/r}$ is strictly decreasing. Hence, for $q_1 \neq q_2$ and $0 < \alpha < 1$, $\beta = 1 - \alpha$ we have

$$\begin{aligned} U_r(\alpha q_1 + \beta q_2, p))^{-r} &= \int \left(\frac{\alpha q_1(x) + \beta q_2(x)}{p(x)} \right)^{1+r} p(x)\, dx \\ &< \alpha \int \left(\frac{q_1(x)}{p(x)} \right)^{1+r} p(x)\, dx + \beta \int \left(\frac{q_2(x)}{p(x)} \right)^{1+r} p(x)\, dx \\ &= \alpha U_r(q_1, p)^{-r} + \beta U_r(q_2, p)^{-r} \\ &= U^{-r} \end{aligned} \tag{29}$$

and since $x^{-1/r}$ is decreasing we obtain

$$U(\alpha q_1 + \beta q_2, p) > U$$

When $-1 < r < 0$, $f(x) = x^{1+r}$ is concave, and one can show by an argument analogous to that leading to (29) that

$$U_r(\alpha q_1 + \beta q_2, p)^{-r} > U^{-r}$$

But in this case $x^{-1/r}$ is increasing, so we obtain the same result (28).

To show that axiom 3 (system independence) is fulfilled, we observe that for $p(x, y) = p_1(x)p_2(y)$ and $q(x, y) = q_1(x)q_2(y)$ one has

$$U_r(q, p) = U_r(q_1, p_1) U_r(q_2, p_2) \tag{30}$$

Thus the maximum of the left-hand side under constraints which affect the two factors on the right hand side separately, is attained exactly when these factors themselves are maximized.

Finally, to prove subset independence, let us put $p(x) = \sum_i \alpha_i p_i(x)$ and $q(x) = \sum_i \beta_i q_i(x)$, where $p_i(x), q_i(x)$ are conditional probability densities on a disjoint partition $(A_1, \ldots, A_n)$ of S, i.e. $p_i(x) = p(x|A_i)$, $q_i(x) = q(x|A_i)$, and $\alpha_i = P(A_i)$, $\beta_i = Q(A_i)$. It follows that

$$\begin{aligned} U_r(q, p)^{-r} &= \sum_i \left(\frac{\beta_i}{\alpha_i} \right)^{1+r} \alpha_i \int \left(\frac{q_i(x)}{p_i(x)} \right)^{1+r} p_i(x) dx \\ &= \sum_i U_r^{-r}(q_i, p_i) \left(\frac{\beta_i}{\alpha_i} \right)^{1+r} \alpha_i \end{aligned} \tag{31}$$

Thus, $U_r(q, p)$ depends only on the coefficients α_i ,and β_i and $U_r(q_i, p_i)$ for $i = 1, \ldots n$. Now in the axiom of strong subset independence (cf. footnote 10), β_i are assumed to be fixed,

and constraint sets $\mathcal{I}_i$ are given for $q_i(x)$. It is demanded that the procedure or maximizing $U(q,p)$ under the constraint

$$I_1 \wedge \cdots \wedge I_n \ : \ \Big(q^\dagger(x) \in \{q(x) \ : \ q(x|A_i) \in \mathcal{I}_i \text{ for } \ i = 1, \ldots n\}\Big)$$

must lead to the same result as maximizing $U(q_i, p_i)$ on the constraint $I_i : (q_i^\dagger(x) \in \mathcal{I}_i$ separately. But it is clear from (31) that $U_r(q,p)$ reaches an extremum only in case all the terms $U_r(q_i, p_i)$ in the sum on the right-hand side are extreme under the given constraints, since they are all non-negative. (There is one exception for this statement, viz. when one of the $U_r(q_i, p_i)$ becomes zero. This can only happen if the measure Q is not absolutely continuous with respect to P, i.e. when $Q(A) \neq 0$ for some set A with $P(A) = 0$. This case is excluded also in Shore and Johnson's proof.)

Next we show the 'only if' part of the assertion. We first employ a theorem of Shore and Johnson which shows that any inference rule which is assumed to maximize some functional $F(q,p)$ and obeys axiom 1, 2 and 5 must be equivalent to the the maximization or minimization of

$$\int h\left(\frac{q(x)}{p(x)}\right) q(x)dx$$

where h is a yet undetermined function. It is slightly more convenient to put $h(y) := y\phi(y)$ so that the above expression reads:

$$\int \phi\left(\frac{q(x)}{p(x)}\right) p(x)dx \tag{32}$$

It is also assumed by Shore and Johnson that the maximization problem can be solved by the Lagrange multiplier technique, from which we may infer that $\phi(y)$ is a continuous and smooth function for $y > 0$. The maximization problem under the constraint

$$\int q(x)f(x)dx = \alpha$$

leads, by the multiplier technique, to

$$\lambda + \mu a(x) + \dot{\phi}\left(\frac{q(x)}{p(x)}\right) = 0$$

where λ and μ are as yet undetermined. In order to obtain a unique solution for q we must, of course, assume that $\dot{\phi}$ is invertible. If $\dot{\phi}$ is continuous, this implies that $\dot{\phi}$ is monotonously increasing or decreasing, so that ϕ is either convex or concave. It can be shown that we have a minimization problem in the first case and a maximization problem in the second case. But these two problems are equivalent, and we can restrict ourselves to the case where ϕ is convex. We now show that if the minimization of (32) obeys system independence for convex ϕ, then $\phi(y) = ay^{1+r} + b$ for some constants a, b.

System independence means that the following two problems should have identical solutions: Let x, y denote possible states of two systems and

(i) Determine the distribution $q(x,y) = q_1(x)q_2(y)$ which minimizes

$$\int \phi\left(\frac{q_1(x)q_2(y)}{p_1(x)p_2(y)}\right) p_1(x)p_2(y)\, dxdy \tag{33}$$

under the constraints $q_1 \in \mathcal{I}_1$, $q_2 \in \mathcal{I}_2$.

(ii) Determine the distributions $q_1(x), q_2(y)$ which minimize

$$\int \phi\left(\frac{q_1(x)}{p_1(x)}\right) dx \quad \int \phi\left(\frac{q_2(y)}{p_2(y)}\right) dy$$

under the same constraints.

Let us focus on system 1. Since p_2 and $\mathcal{I}_2$ can be chosen arbitrarily, a necessary condition for the validity of system independence is that the following holds: The distribution q_1 that minimizes (33) for any q_2, p_2 also minimizes

$$\int \phi\left(\frac{q_1(x)}{p_1(x)}\right) p_1(x) dx$$

and conversely. Now consider a case with only three possible outcomes: $S = \{x_1, x_2, x_3\}$ and let $q_2(x_1)/p_2(x_1) = q_2(x_2)/p_2(x_2) = q_2(x_3)/p_2(x_3) = \alpha$. Put $q_1(x_i)/p_1(x_i) = \beta_i$, $i = 1, 2, 3$, and let $\mathcal{I}_1$ consists of a one-parameter family of distributions $\{q_{1\theta} : \theta \in I\!R\}$ so that we can parametrize: $q_{1\theta}(x_i) = p_1(x_i)\beta_i(\theta)$.

The solution of problem (i) is now determined by the condition:

$$\sum_i \dot{\phi}(\alpha\beta_i)\frac{d\beta}{d\theta} = 0$$

and for problem (ii):

$$\sum_i \dot{\phi}(\beta_i)\frac{d\beta_i}{d\theta} = 0.$$

Further, the normalization $\sum_i q_1(x_i) = 1$ implies:

$$\frac{d}{d\theta}\sum_i q_1(x_i) = \sum_i p_1(x_i)\frac{d\beta_i}{d\theta} = 0$$

In particular, consider the case $p_i = 1/3$, and $d\beta/d\theta = (\frac{1}{2}, \frac{1}{2}, -1)$. The condition of the equivalence of problems (i) and (ii) is then:

$$\frac{1}{2}(\dot{\phi}(\beta_1) + \dot{\phi}(\beta_2)) = \dot{\phi}(\beta_3) \Longleftrightarrow \frac{1}{2}(\dot{\phi}(\alpha\beta_1) + \dot{\phi}(\alpha\beta_2)) = \dot{\phi}(\alpha\beta_3)$$

Eliminating β_3 from these equations, and writing:

$$M_{\dot{\phi}}(\beta_1, \beta_2) := \dot{\phi}^{-1}(\dot{\phi}(\beta_1 + \dot{\phi}(\beta_2)/2)$$

we obtain the condition:

$$M_{\dot{\phi}}(\alpha\beta_1, \alpha\beta_2) = \alpha M_{\dot{\phi}}(\beta_1, \beta_2)$$

I.e., the expression $M_{\dot{\phi}}$ must be linear. A theorem of Hardy, Littlewood and Polya (1934, p. 68) shows that any such U for a monotonous $\dot{\phi}$, must be equivalent to the choice $\dot{\phi} = x^r$ for some $r \in I\!R$. (Equivalence means: equality upto multiplicative and additive constants which have no effect on the value of M_ϕ.) Hence ϕ is equivalent to the choice $\phi(x) = x^{1+r}$, and the procedure is equivalent to either a minimization of

$$\int (\frac{q(x)}{p(x)}^{1+r} p(x)\, dx$$

if ϕ is convex, i.e. $r > 0$ or $r < -1$; or to a maximization of this expression if ϕ is concave, for $-1 < r < 0$. QED. The further restriction to $r > -1$ is explained in the final paragraph of section 6.

References

Aczél, J. and Daróczy, Z. (1975), *On Measures of Information and their Characterizations* (New York: Academic Press).

Balian, R. (1991), *From Microphysics to Macrophysics* (Berlin: Springer).

Boole, G. (1862), 'On the Theory of Probability', in *Collected Logical works*, Vol I (Lasalle, Illinois: Open Court, 1952).

Bhattacharyya, A. (1943), 'On a Measure of Divergence between Two Statistical Populations Defined by their Probability Distributions' *Bulletin of the Calcutta Mathematical Society* **35**, 99-109.

Buck, B. and Macaulay, V.A. (1991), *Maximum Entropy in Action*, (Oxford: Clarendon Press).

Callen, H. (1960), *Thermodynamics* (New York: John Wiley).

Csiszár, I. (1985), 'An Extended Maximum Entropy Principle and a Bayesian Justification', in J.M. Bernardo et al. (eds), *Bayesian Statistics 2*, (Amsterdam: North-Holland and Valencia University Press), pp. 83–98.

Denbigh K.G., and Denbigh, J.S. (1985), *Entropy in its relation to incomplete knowledge* (Cambridge: Cambridge University Press).

Dias, P.M. and Shimony, A. (1981), 'A Critique of Jaynes' Maximum Entropy Principle', *Advances in Applied Mathematics* **2**, 172–211.

Dougherty, J.P. (1993), 'Explaining Statistical Mechanics', *Studies in History and Philosophy of Modern Physics* **24**, 843–866.

Edwards, A.W.F. (1972), *Likelihood*, (Cambridge: Cambridge University Press; expanded edition John Hopkins University Press, 1992).

Ellis, R.L. (1850), 'Remarks on an Alleged Proof of the Method of Least Squares, Contained in a Late Number of the *Edinburgh Review*', in W. Walton (ed), *Mathematical and other Writings of R.L. Ellis* (Cambridge: Cambridge University Press, 1863), pp. 53–61.

Faddeev, D.K. (1957), 'Zum Begriff der Entropie eines endliches Wahrscheinlichkeitsschemas', in H. Grell (ed), *Arbeiten zur Informationstheorie,* Vol. 1 (Berlin: Deutscher Verlag der Wissenschaften,), pp. 88–91. Russian original in *Uspekhi Matematicheskikh Nauk* **11**, (1956), 227–231.

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

Fraassen, B.C. van (1981), 'A Problem for Relative Information Minimizers in Probability Kinematics', *British Journal for the Philosophy of Science* **32**, 375.

Fraassen, B.C. van, Hughes R.I.G., and Harman, G. (1986), 'A Problem for Relative Information Minimizers, Continued', *British Journal for the Philosophy of Science*, **37**, 453–463.

Fraassen, B.C. van (1989), *Laws and Symmetry*, (Oxford: Clarendon Press).

Friedman, K. and Shimony, A. (1971), 'Jaynes's Maximum Entropy Prescription and Probability Theory', *Journal of Statistical Physics* **3**, 381–384.

Gelfand, I.M., Yaglom A.M. and Kolmogorov A.N. (1958) , 'Zur algemeinen Definition der Infomation', in H. Grell (ed), *Arbeiten zur Informationstheorie,* Vol. 2 (Berlin: Deutscher Verlag der Wissenschaften. Russian original in *Doklady Akademii Nauk* **111**, (1956) 745–748.

Hacking, I. (1971), 'Equipossibility Theories of Probability', *British Journal for the Philosophy of Science* **22**, 339–355.

Hacking, I. (1975), *The Emergence of Probability* (Cambridge: Cambridge University Press).

Hardy, G., Littlewood J.E. and Pólya, G. (1934), *Inequalities* (Cambridge: Cambridge University Press).

Hilgevoord J. and Uffink, J. (1991), 'Uncertainty in Prediction and in Inference', *Foundations of Physics* **21**, 323.

Hobson, A. (1971), *Concepts in Statistical Mechanics* (New York: Gordon and Breach).

Jaynes, E.T. (1957a), 'Information Theory and Statistical Mechanics I', *Physical Review* **106**, 620–630. Reprinted in Jaynes (1981), pp. 6–16.

Jaynes, E.T. (1957b), 'Information Theory and Statistical Mechanics II', *Physical Review* **108**, 171–190. Reprinted in Jaynes (1981), pp. 19–38.

Jaynes, E.T. (1968), 'Prior Probabilities', *IEEE Transactions on Systems Science and Cybernetics* **SSC-4** 227–241. Reprinted in Jaynes (1981) pp. 116–130.

Jaynes, E.T. (1973), 'The Well-Posed Problem', *Foundations of Physics* **3**, 477–493. Reprinted in Jaynes (1981), pp. 133–147.

Jaynes, E.T. (1978), 'Where do we stand on maximum entropy?', in R.D. Levine and M. Tribus (eds), *the Maximum Entropy Formalism* (Cambridge Massachussetts: MIT Press). Reprinted in Jaynes 1981), pp. 210–314.

Jaynes, E.T. (1981), *Papers on Probability, Statistics and Statistical Physics* R. Rosenkrantz (ed) (Dordrecht: Reidel).

Jaynes, E.T. (1985), 'Some Random Observations', *Synthese* **63**, 115–138.

Jaynes, E.T. (1986), 'Monkeys, Kangaroos and N', in J.H. Justice (ed), *Maximum Entropy and Bayesian Methods in Applied Statistics* (Cambridge: Cambridge University Press), pp. 27- -57.

Johnson R.W. and Shore, J.E. (1983), 'Comments on and Correction to "Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-Entropy"', *IEEE Transactions on Information Theory* **IT-29**, 942–943.

Johnson, R.W. and Shore J.E. (1984), 'Comment on "Consistent Inference of Probabilities for Reproducible Experiments"', *Physical Review Letters* **55**, 336.

Keynes, J.M. (1973), *Treatise on Probability* (London: The Macmillan Press).

Khinchin, A.I. (1957), *Mathematical Foundations of Information Theory* (New York: Dover).

Kolmogorov, A.N. (1957), 'Theorie der Nachrichtenübermittlung', in H. Grell (ed), Arbeiten zur Informationstheorie, Vol. 1. (Berlin: Deutscher Verlag der Wissenschaften).

Kries, J. von (1927), *Die Principien der Wahrscheinlichkeitsrechnung* (Tübingen: Verlag von J.C.B. Mohr, 2nd edition).

Kries, J. von (1916), *Logik* (Tübingen: Verlag von J.C.B. Mohr).

Kullback, S. (1959), *Information Theory and Statistics* (New York: Wiley).

Laplace P.S. de (1829), *Essai philosophique sur les Probabilités* (Bruxelles: H. Remy, 5th edition).

Lavis D.A. and Milligan, P.J. (1985), 'The Work of E.T. Jaynes on Probability, Statistics and Statistical Physics', *British Journal for the Philosophy of Science* **36**, 193–210.

Mises, R. von (1928), *Wahrscheinlichkeit, Statistik und Wahrheit* (Wien: Springer).

Penrose, O. (1979), 'Foundations of Statistical Mechanics', *Reports on the Progress of Physics* **42**, 1937–2006.

Reichenbach, H. (1935), *Wahrscheinlichkeitslehre* (Leiden: A.W. Sijthoff).

Rényi, A. (1962), *Wahrscheinlichkeitsrechnung* (Berlin: Deutscher Verlag der Wissenschaften).

Roberts, A.W. and Varberg, D.E. (1973), *Convex Functions* (New York: Academic Press).

Seidenfeld, T. (1979), 'Why I am not an Objective Bayesian', *Theory and Decision* **11**, 413–440.

Seidenfeld, T. (1986), 'Entropy and Uncertainty' *Philosophy of Science* **53**, 467–491.

Shannon, C.E. (1948), 'A Mathematical Theory of Communication', *Bell Systems Technical Journal* **27**, 379–423 and 623–656.

Shimony, A. (1984), 'Comment on "Consistent Inference of Probabilities for Reproducible Experiments"', *Physical Review Letters* **55**, 1030.

Shimony, A. (1985), 'The Status of the Principle of Maximum Entropy', *Synthese* **63**, 35–53.

Shore J.E. and Johnson, R.W. (1980), 'Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-Entropy', *IEEE Transactions on Information Theory* **IT-26**, 26–37.

Shore, J.E. and Johnson, R.W. (1981), 'Properties of Cross-Entropy Minimization', *IEEE Transactions on Information Theory* **IT-27**, 472–482.

Skilling, J. (1984a), 'The Maximum Entropy Method', *Nature* **309**, 748–749.

Skilling, J. (1984b), 'Reply to a letter by Titterington', *Nature* **312**, 382.

Skilling, J. (1988), 'The axioms of maximum entropy', in G.J. Erickson and C.R. Smith (eds), *Maximum Entropy and Bayesian Methods in Science and Engineering*, Vol.1 (Dordrecht: Kluwer), pp. 173–187.

Skilling, J. (1989), 'Classical maximum entropy', in J. Skilling (ed), *Maximum Entropy and Bayesian Methods* (Dordrecht: Kluwer), pp. 45–52.

Tikochinsky, Y., Tishby N.Z. and Levine, R.D. (1984a), 'Consistent Inference of Probabilities for Reproducible Experiments', *Physical Review Letters* **52**, 1357–1360.

Tikochinsky, Y., Tishby N.Z. and Levine, R.D. (1984b), 'Alternative Approach to Maximum-Entropy Inference' *Physical Review A* **30**, 2638–2644.

Tikochinsky, Y., Tishby, N.Z. and Levine R.D. (1984c), [Response to Johnson and Shore, 1984] *Physical Review Letters* **55**, 337.

Tikochinsky, Y., Tishby N.Z. and Levine, R.D. (1984d), [Response to Shimony, 1984] *Physical Review Letters* **55**, 1031.

Tisza, L. (1966), *Generalized Thermodynamics*, (Cambridge Massachusetts: MIT Press).

Uffink, J. (1990), *Measures of Uncertainty and the Uncertainty Principle*, (Utrecht: Utrecht University).

Uffink, J. (1995), *Constraints in the Maximum Entropy Principle* (preprint).

Williams, P.M. (1980), 'Bayesian Conditionalisation and the Principle of Minimum Information', *British Journal for the Philosophy of Science* **31**, 131–144.

Wootters, W.K. (1981), 'Statistical Distance in Hilbert Space', *Physical Review D* **23**, 357–362.