The Principle of Maximum Entropy and a Problem in Probability Kinematics

1. Introduction

There are cases in which we cannot use Bayes' formula to deliver a set of updated probabilities given a set of probabilities together with a new observation. The Bayesian method presupposes that evidence comes in the form of an event. As Jeffrey observes, however, the evidence may not relate the certainty of an event but a reassessment of its uncertainty or its probabilistic relation to other events (see Jeffrey, 1965, 153ff), expressible in a shift in expectation (see Hobson, 1971). Bas van Fraassen has come up with a good example from the 1980 comedy film *Private Benjamin* (see van Fraassen 1981), in which Goldie Hawn portrays a Jewish-American woman (Judy Benjamin) who joins the U.S. Army.

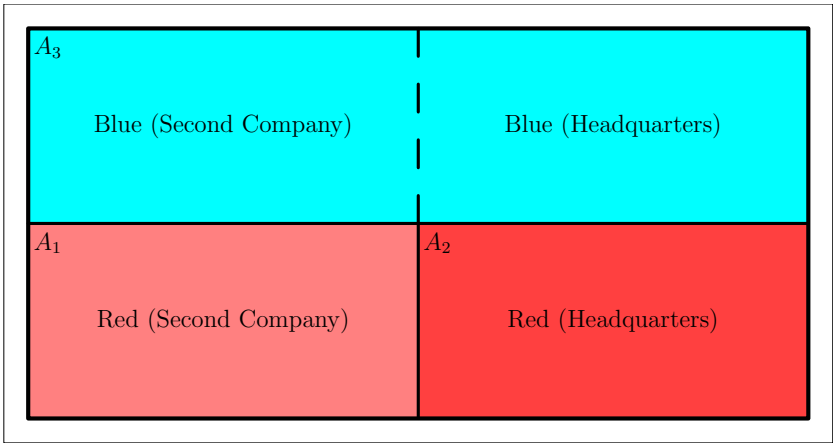

Figure 1: Judy Benjamin's map. Blue territory (A_3) is friendly and does not need to be divided into a Headquarters and a Second Company area.

In the movie, Judy Benjamin is on an assignment and lands in a place

where she is not sure of her location. She is on team Blue. Because of the map, her probability of being in Blue territory equals the probability of being in Red territory, and being in the Red Second Company area equals the probability of being in the Red Headquarters area. Her commanders inform Judy by radio that in case she is in Red territory, her chance of being in the Red Headquarters area is three times the chance of being in the Red Second Company area.

At the heart of our investigation are two incompatible but independently plausible intuitions regarding Judy's choice of updated probabilities for her location (much more in the next section). First we note, however, that there is no immediately obvious event space in which we can condition on an event E. Grove and Halpern (1997) have written an article on how to construct such event spaces and then condition on the event that Judy Benjamin receives the information that she receives from her commanders. They admit, however, that the construction of such spaces (sometimes called retrospective conditioning) is an exercise in filling in missing details and supplying information not contained in the original problem.

Joseph Halpern writes in *Reasoning About Uncertainty* that "there is no escaping the need to understand the details of the application" (Halpern, 2003, 423) and concludes that the principle of maximum entropy (from now on MAXENT for abbreviation) is a valuable tool, but should be used with care (see Grove and Halpern, 1997, 110), explicitly basing his remark on the counterintuitive behaviour of the Judy Benjamin problem. Diaconis and

Zabell state "that any claims to the effect that maximum-entropy revision is the only correct route to probability revision should be viewed with considerable caution" (Diaconis and Zabell, 1982, 829). "Great caution" (1994, 456) is also what Colin Howson and Allan Franklin advise about the claim that the updated probabilities provided by MAXENT are as like the original probabilities as it is possible to be given the constraints imposed by the data.

Igor Douven and Jan-Willem Romeijn agree with Richard Bradley that "even Bayes' rule 'should not be thought of as a universal and mechanical rule of updating, but as a technique to be applied in the right circumstances, as a tool in what Jeffrey terms *the art of judgment*.' In the same way, determining and adapting the weights [epistemic entrenchment] supposes, or deciding when Adams conditioning applies, may be an art, or a skill, rather than a matter of calculation or derivation from more fundamental epistemic principles" (Douven and Romeijn, 2009, 16) (for the Bradley quote see Bradley, 2005, 362).

Teddy Seidenfeld gives his own reasons why he is not an objective Bayesian in an article entitled "Why I Am Not an Objective Bayesian" (1979) (claiming that in particular circumstances involving noise factors MAXENT will inappropriately provide more information based on less evidence). Jos Uffink targets especially Shore and Johnson's assumptions when they identify MAXENT as the unique method of determining updated probability distributions, given certain types of constraints. Uffink shows

how a more reasonable restatement of Shore and Johnson's assumptions results in a whole class of updating procedures, the so-called Rényi entropies (see Uffink, 1995). This also appears to be van Fraassen's conclusion when he suggests that MAXENT is a special instance of a family of principles which are consistent relative to specified assumptions (see van Fraassen et al. 1986). Dias and Shimony provide a very interesting case of failure for MAXENT that we will not address in this paper because it is not relevant to the Judy Benjamin problem (see Dias and Shimony, 1981), although we hope to address it at a later date.

What is lacking in the literature is an explanation by MAXENT advocates of the counterintuitive behaviour of the cases repeatedly quoted by their adversaries. This is especially surprising as we are not dealing with an array of counter-examples but only a handful, the Judy Benjamin problem being prime among them. In Halpern's textbook, for example, the reasoning is as follows: MAXENT is a promising candidate which delivers unique updated probability distributions; but, unfortunately, there is counterintuitive behaviour in one specific case, the Judy Benjamin case (see Halpern, 2003, 110, 119); therefore, we must abide by the eclectic principle of considering not only MAXENT, but also lower and upper probabilities, Dempster-Shafer belief functions, possibility measures, ranking functions, relative likelihoods, and so forth. The human inquirer is the final arbiter between these conditionalization methods.

We will undermine the notion that MAXENT's solution for the Judy

Benjamin problem is counterintuitive. The intuition that MAXENT's solution for the Judy Benjamin problem violates (call it T1) is based on fallacious independence and uniformity assumptions. There is another powerful intuition (call it T2) in direct contradiction to T1 which MAXENT obeys. Therefore, Halpern does not give us sufficient grounds for the eclecticism advocated throughout his book. We will show that another intuitive approach, the powerset approach, lends significant support to the solution provided by MAXENT for the Judy Benjamin problem, especially in comparison to intuition T1, many of whose independence and uniformity assumptions it shares.

Dealing in blatant stereotypes for a moment, there is a long-standing disagreement between philosophers on the one hand and physicists on the other hand whether (phil) updating probabilities is irreducibly accompanied by thoughtful deliberation choosing between approaches depending on individual problems, or (phys) problems are ill-posed if they do not contain the information necessary to let a non-arbitrary, non-subjective process arrive at a unique updated probability assessment. In the literature, Judy Benjamin serves as an example to defend in favour of the philosophers what I shall call the full employment theorem of probability kinematics.

The full employment theorem of probability kinematics claims that MAXENT is only one of many different strategies to update probabilities. In order to decide which strategy is the most appropriate for your problem you need a resident formal epistemologist to do the thinking and weigh the

intuitions for you. For a fee, of course. Thus formal epistemologists will always be fully employed. (E.T. Jaynes makes similar observations when he derisively talks about the statistician-client relationship as one between a doctor and his patient, see Jaynes and Bretthorst, 1998, 492 and 506.) There is an analogous full employment theory in computer science about writing computer programs which has been formally proved to be true. Our contention is that no such proof is forthcoming in probability kinematics. The case rests in a significant measure (see Halpern's book) on a counterexample to MAXENT, the Judy Benjamin problem. We show in this article that our intuitions are initially misguided about the Judy Benjamin problem because we make independence and uniformity assumptions that on closer examination are not tenable.

2. Two Intuitions

There are two pieces of information relevant to Judy Benjamin when she decides on her updated probability assignment. We will call them (MAP) and (HDQ). As in figure 1, A_1 is the Red Second Company area, A_2 is the Red Headquarters area, A_3 is Blue territory. Judy presumably wants to be in Blue territory, but if she is in Red territory, she would prefer their Second Company area (where enemy soldiers are not as well-trained as in the Headquarters area).

(MAP) Judy has no idea where she is. She is on team Blue. Because of the map, her probability of being in Blue territory equals the probability

of being in Red territory, and being in the Red Second Company area equals the probability of being in the Red Headquarters area.

(HDQ) Her commanders inform Judy that in case she is in Red territory, her chance of being in their Headquarters area is three times the chance of being in their Second Company area.

In formal terms (sloppily writing A_i for the event of Judy being in A_i),

$$2 \cdot P(A_1) = 2 \cdot P(A_2) = P(A_3) \qquad \text{(MAP)}$$

$$q = P(A_2|A_1 \cup A_2) = \frac{3}{4} \qquad \text{(HDQ)}$$

(HDQ) is partial information because in contrast to the kind of evidence we are used to in Bayes' formula (such as 'an even number was rolled'), and to the kind of evidence needed for Jeffrey's rule (where a partition of the whole event space and its probability reassignment is required, not only $A_1 \cup A_2$, but see here the objections in Douven and Romeijn, 2009), the scenario suggests that Bayesian conditionalization and Jeffrey's rule are inapplicable. We are interested in the most defensible updated probability assignment(s) and will express them in the form of a normalized odds vector (q_1, q_2, q_3), following van Fraassen (1981). q_i is the updated probability $Q(A_i)$ that Judy Benjamin is in A_i. Let P be the probability distribution prior to the new obseration and p_i the individual 'prior' probabilities. These probabilities are not to be confused with prior

probabilities that precede any kind of information. In the spirit of probability update, or probability kinematics, we will for the rest of the article refer to prior probabilities as probabilities prior to an observation and the subsequent update. The q_i sum to 1 (this differs from van Fraassen's canonical odds vector, which is proportional to the normalized odds vector but has 1 as its first element). We define

$$t = \frac{q}{1-q}$$

t is the factor by which (HDQ) indicates that Judy's chance of being in A_2 is greater than being in A_1. In Judy's particular case, $t = 3$ and $q = 0.75$. Van Fraassen found out with various audiences that they have the following intuition:

T1 (HDQ) does not refer to Blue territory and should not affect $P(A_3)$: $q_3 = p_3 (= 0.50)$.

There is another, conflicting intuition (due to Peter Williams via personal communication with van Fraassen, see van Fraassen 1981, 379):

T2 If the value of q approaches 1 (in other words, t approaches infinity) then q_3 should approach 2/3 as the problem reduces to one of ordinary conditioning. (HDQ) would turn into 'if you are in Red territory you are almost certainly in the Red Headquarters area.' Considering (MAP), q_3 should approach 2/3. Continuity considerations pose a contradiction to T1.

To parse these conflicting intuitions, we will introduce several methods to provide G, the function that maps q to the appropriate normalized updated odds vector (q_1, q_2, q_3). The first method is extremely simple and accords with intuition T1: $G_{\text{ind}}(q) = (0.5(1-q), 0.5q, 0.5)$. In Judy's particular case with $t = 3$ the normalized odds vector is (ind stands for independent):

$$G_{\text{ind}}(0.75) = (0.125, 0.375, 0.500)$$

Both Grove and Halpern (1997) as well as Douven and Romeijn (2009) make a case for this distribution, Grove and Halpern by using Bayesian conditionalization on the event of the message being transmitted to Judy, Douven and Romeijn by using Jeffrey's rule (because they believe that T1 is in this case so strong that $Q(A_3) = P(A_3)$ is as much of a constraint as (MAP) and (HDQ), yielding a Jeffrey partition). T1, however, conflicts with the symmetry requirements outlined in van Fraassen et. al. (1986).

Van Fraassen introduces various updating methods which do not conflict with those symmetry requirements, the most notable of which is MAXENT. Shore and Johnson have already shown that, given certain assumptions (which have been heavily criticized, however, e.g. Uffink, 1996), MAXENT produces a unique updated probability assignment. The minimum information discrimination theorem of Kullback and Leibler (see, for example, Csiszár, 1967, section 3) demonstrates how Shannon's entropy and the Kullback-Leibler Divergence formula can provide the least informative

updated probability assignment (with reference to the prior probability assignment) obeying the constraint posed by the evidence. We show how this is done in Appendix I. The idea is to define a space of probability distributions, make sure that the constraint identifies a closed, convex subset in this space, and then determine which of the distributions in the closed, convex subset is least distant from the prior probability distribution in terms of information (using the minimum information discrimination theorem). It is necessary for the uniqueness of this least distant distribution that the subset be closed and convex (in other words, that the constraints be affine, see Csiszár, 1967).

For Judy Benjamin, MAXENT suggests the following normalized odds vector:

$$G_{\max}(0.75) \approx (0.117, 0.350, 0.533) \tag{1}$$

The updated probability of being on Blue territory (A_3) has increased from 50% to approximately 53%. Grove and Halpern find this result "highly counterintuitive" (Grove and Halpern, 1997, 2). Van Fraassen summarizes the worry:

> It is hard not to speculate that the dangerous implications of being in the enemy's Headquarters area are causing Judy Benjamin to indulge in wishful thinking, her indulgence becoming stronger as her conditional estimate of the danger increases. (Van Fraassen, 1981, 379)

There are two ways in which we can arrive at result (1).

We may either use Jaynes' constraint rule and find the updated probability distribution that is both least informative with respect to Shannon's entropy and in accordance with the constraint (using Dempster's Rule of Combination, which together with the constraint rule is equivalent to the principle of minimum cross-entropy, see Cover and Thomas, 2006, 409, exercise 12.2.), or use the Kullback-Leibler Divergence and differentiate it to obtain where it is minimal.

The constraint rule has the advantage of providing results when the derivative of the Kullback-Leibler Divergence is difficult to find. This not being the case for Judy, we go the easier route of the second method and provide a more general justification for the constraint rule in Appendix I, together with its application to the Judy Benjamin case.

The Kullback-Leibler Divergence is

$$D(Q, P) = \sum_{i=1}^{m} q_i \log \frac{q_i}{p_i}$$

We fill in the explicit details from Judy Benjamin's situation and differentiate the expression to obtain the minimum (by setting the derivative to 0).

$$\frac{\partial}{\partial q_1}(q_1 \log_2(4q_1) + tq_1 \log_2(4tq_1) + (1 - (t+1)q_1) \log_2 2(1 - (t+1)q_1)) = 0$$

The resulting expression for $G_{\max}$ is

$$G_{\max}(q) = \left(\frac{C}{1+Ct+C}, t\frac{C}{1+Ct+C}, 1-(t+1)\frac{C}{1+Ct+C} \right)$$

where

$$C = 2^{-\frac{t\log_2 t+t+1}{1+t}}$$

Figures 2 and 3 show in diagram form the distribution of (q_1, q_2, q_3) depending on the value of q (between 0 and 1), respectively following intuition T1 and MAXENT. Notice that in accordance with intuition T2, MAXENT provides a result where $q_3 \to 2/3$ for q approaching 0 or 1.

3. Epistemic Entrenchment and Coarsening at Random

Even though T1 is an understandably strong intuition, it does not take into account that the information given to Judy by her commanders may be dependent on whether she is in Blue or in Red territory. To underline this objection to intuition T1 we want to consider three scenarios, any of which may form the basis of the partial information provided by her commanders.

I Judy was dropped off by a pilot who flipped two coins. If the first coin landed H, then Judy was dropped off in Blue territory, otherwise in Red territory. If the second coin landed H, she was dropped off in the Headquarters area, otherwise in the Second Company area. Judy's commanders find out that the second coin was biased $q : 1-q$ toward H with $q = 0.75$. The normalized odds vector is

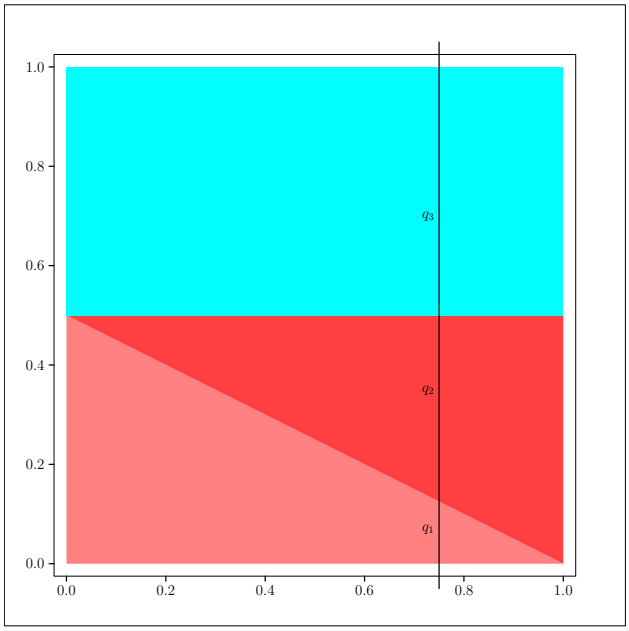

Figure 2: Judy Benjamin's updated probability assignment according to intuition T1. $0 < q < 1$ forms the horizontal axis, the vertical axis shows the updated probability distribution (or the normalized odds vector) (q_1, q_2, q_3). The vertical line at $q = 0.75$ shows the specific updated probability distribution $G_{\text{ind}}(0.75)$ for the Judy Benjamin problem.

$G_{\text{I}}(0.75) = (0.125, 0.375, 0.500)$ and agrees with T1, because the choice of Blue or Red is completely independent from the choice of the Red Headquarters area or the Red Second Company area.

II The pilot randomly lands in any of the four quadrants and rolls a die. If she rolls an even number, she drops off Judy. If not, she takes her to another (or the same, the choice happens with replacement) randomly

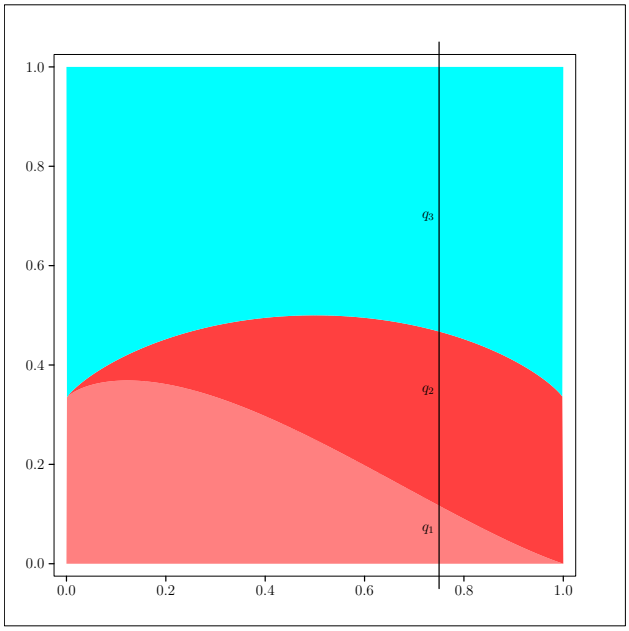

Figure 3: Judy Benjamin's updated probability assignment using MAXENT. $0 < q < 1$ forms the horizontal axis, the vertical axis shows the updated probability distribution (or the normalized odds vector) (q_1, q_2, q_3). The vertical line at $q = 0.75$ shows the specific updated probability distribution $G_{\max}(0.75)$ for the Judy Benjamin problem.

selected quadrant to repeat the procedure. Judy's commanders find out, however, that for A_1, the pilot requires a six to drop off Judy, not just an even number. The normalized odds vector in this scenario is $G_{\text{II}}(0.75) = (0.1, 0.3, 0.6)$ and does not accord with T1.

III Judy's commanders have divided the map into 24 congruent rectangles, A_3 into twelve, and A_1 and A_2 into six rectangles each (see

figures 4 and 5). They have information that the only subsets of the 24 rectangles in which Judy Benjamin may be located are such that they contain three times as many A_2 rectangles than A_1 rectangles. The normalized odds vector in this scenario is $G_{\text{III}}(0.75) \approx (.108, .324, .568)$ (evaluating almost 17 million subsets).

I–III demonstrate the contrast between scenarios when independence is true and when it is not. Douven and Romeijn's capital mistake in their paper is that they assume that the Judy Benjamin problem is analogous to their example of Sarah and the sundowners at the Westcliff (see Douven and Romeijn, 2009, 7). Sarah, however, knows that whether it rains or not is independent of her activity the next night, whereas in Judy Benjamin we have no evidence of such independence, as scenario II demonstrates. Douven and Romeijn's reliance on intuition T1 leads them to apply Jeffrey's rule to the Judy Benjamin problem with the additional constraint $Q(A_3) = P(A_3)$. They claim that in most cases "the learning of a conditional is or would be irrelevant to one's degree of belief for the conditional's antecedent ... the learning of the relevant conditional should intuitively leave the probability of the antecedent unaltered" (Douven and Romeijn, 2009, 9). This, according to Douven and Romeijn, is the usual epistemic entrenchment and applies in full force to the Judy Benjamin problem. They give an example where the epistemic entrenchment could go the other way and leave the consequent rather than the antecedent unaltered (Kate and Henry, see Douven and Romeijn, 2009, 13).

Another at first blush forceful argument that MAXENT's solution for the Judy Benjamin problem is counterintuitive has to do with coarsening at random, or CAR for short. It is spelled out in Grünwald and Halpern (2003). Grünwald and Halpern see a parallel between the Judy Benjamin problem and Martin Gardner's Three Prisoners problem (see Gardner, 1959, 180f). In the Three Prisoners problem, three men (A, B, and C) are under sentence of death when the governor decides to pardon one of them. The warden of the prison knows which of the three men is pardoned, but none of the men do. In a private conversation, A says to the warden, Tell me the name of one of the others who will be executed—it will not give anything away whether I will be executed or not. The warden agrees and tells A that B will be executed. For the puzzling consequences, see the wealth of literature on the Three Prisoners problem or the Monty Hall problem.

According to Grünwald and Halpern, for problems of this kind (Judy Benjamin, Three Prisoners, Monty Hall) there are naive and sophisticated spaces to which we can apply probability updates. If A uses the naive space, for example, he comes to the following conclusion: of the three possibilities that (A,B), (A,C), or (B,C) are executed, the warden's information excludes (A,C). (A,B) and (B,C) are left over, and because A has no information about which one of these is true his chance of not being executed is 0.5. His chance of survival has increased from one third to one half.

Grünwald and Halpern show, correctly, that the application of the naive space is illegitimate because the CAR condition does not hold. More

generally, Grünwald and Halpern show that updating on the naive space rather than the sophisticated space is legitimate for event type observations always when the set of observations is pairwise disjoint or, when the events are arbitrary, only when the CAR condition holds. For Jeffrey type observations, there is a generalized CAR condition which applies likewise. For affine constraints on which we cannot use Jeffrey conditioning (or, a fortiori, conditional probabilities) MAXENT "essentially never gives the right results" (Grünwald and Halpern, 2003, 243).

Grünwald and Halpern conclude that "working with the naive space, while an attractive approach, is likely to give highly misleading answers" (246), especially in the application of MAXENT to naive spaces as in the Judy Benjamin case "where applying [MAXENT] leads to paradoxical, highly counterintuitive results" (245). For the Three Prisoners problem, Jaynes' constraint rule would supposedly proceed as follows: the vector of prior probabilities for (A,B), (A,C), and (B,C) is $(1/3, 1/3, 1/3)$. The constraint is that the probability of (A,C) is zero, and a simple application of the constraint rule yields $(1/2, 0, 1/2)$ for the vector of updated probabilities. The CAR condition for the naive space does not hold, therefore the result is misleading.

By analogy, using the constraint rule on the naive space for the Judy Benjamin problem yields $(0.117, 0.350, 0.533)$, but as the CAR condition fails in even the simplest settings for affine constraints ("CAR is (roughly speaking) guaranteed *not* to hold except in 'degenerate' situations" (251),

emphasis in the original), it certainly fails for the Judy Benjamin problem, for which constructing a sophisticated space is complicated (see Grove and Halpern, 1997, where the authors attempt such a construction by retrospective conditioning).

The analogy, however, is misguided. The constraint rule has been formally shown to generalize Jeffrey conditioning, which in turn has been shown to generalize standard conditioning (the authors admit as much in Grünwald and Halpern, 2003, 262). We can solve both the Monty Hall problem and the Three Prisoners problem by standard conditioning, not using the naive space, but simply using the correct space for the probability update. For the Three Prisoners problem, for example, the warden will say either 'B' or 'C' in response to A's inquiry. Because A has no information that would privilege either answer the probability that the warden says 'B' and the probability that the warden says 'C' equal each other and therefore equal 0.5. Here is the difference between using the naive space and using the correct space, but either way using standard conditional probabilities:

$$P(\text{'A is pardoned'}|\text{'B will be executed'}) =$$

$$\frac{P(\text{'A is pardoned'})}{P(\text{'A is pardoned'}) + P(\text{'C is pardoned'})} = \frac{1}{2} \text{ (incorrect)}$$

$$P(\text{'A is pardoned'}|\text{'warden says B will be executed'}) =$$

$$\frac{P(\text{`A is pardoned' and `warden says B will be executed'})}{P(\text{`warden says B will be executed'})} = \frac{1/6}{1/2} = \frac{1}{3} \text{ (correct)}$$

Why is the first equation incorrect and the second one correct? Information theory gives us the right answer: in the first equation, we are conditioning on a watered down version of the evidence (watered down in a way that distorts the probabilities because we are not 'coarsening at random'). 'Warden says B will be executed' is sufficient but not necessary for 'B will be executed.' The former proposition is more informative than the latter proposition (its probability is lower). Conditioning on the latter proposition leaves out relevant information contained in the wording of the problem.

Because MAXENT always agrees with standard conditioning, MAXENT gives the correct result for the Three Prisoners problem. For the Judy Benjamin problem, there is no defensible sophisticated space and no watering down of the evidence in what Grünwald and Halpern call the 'naive' space. The analogy between the Three Prisoners problem and the Judy Benjamin problem as it is set up by Grünwald and Halpern fails. A successful criticism would be directed at the construction of the 'naive' space: this is what we just accomplished for the Three Prisoners problem. There is no parallel procedure for the Judy Benjamin problem. The 'naive' space is all we have, and MAXENT is the appropriate tool to deal with this lack of information.

4. The Powerset Approach

In this section, we will focus on scenario III and consider what happens

when the grain of the partition becomes finer. We call this the powerset approach. Two remarks are in order: First, the powerset approach has little independent appeal. The reason behind using MAXENT is that we want our evidence to have just the right influence on our updated probabilities, i.e. neither over-inform nor under-inform. There is no corresponding reason why we should update our probabilities using the powerset approach.

Second, what the powerset approach does is lend support to another approach. In this task, it is persuasive because it tells us what would happen if we were to divide the event space into infinitesimally small, uniformly weighed, and independent 'atomic' bits of information. It would be especially interesting if the powerset approach did not support the independence and uniformity assumptions of intuition T1, because both of these features are strongly represented in the powerset approach. On its own the powerset approach is just what Grünwald and Halpern call a naive space, for which CAR does not hold. Hence the powerset approach will not give us a precise solution for the problem, although it may with some plausibility guide us in the right direction—especially if despite all its independence and uniformity assumptions it significantly disagrees with intuition T1.

Let's assume a partition $\{B_i\}_{i=1,\ldots,4n}$ of $A_1 \cup A_2 \cup A_3$ into sets that are of equal measure μ and whose intersection with A_i is either the empty set or the whole set itself (this is the division into rectangles of scenario III). (MAP) dictates that the number of sets covering A_3 equals the number of

sets covering $A_1 \cup A_2$. For convenience, we assume the number of sets covering A_1 to be n. Let $\mathcal{C}$, a subset of the powerset of $\{B_i\}_{i=1,\ldots,4n}$, be the collection of sets which agree with the constraint imposed by (HDQ), i.e.

$$C \in \mathcal{C} \text{ iff } C = \{C_j\} \text{ and } t\mu\left(\bigcup C_j \cap A_1\right) = \mu\left(\bigcup C_j \cap A_2\right)$$

In figures 4 and 5 there are diagrams of two elements of the powerset of $\{B_i\}_{i=1,\ldots,4n}$. One of them (figure 4) is not a member of $\mathcal{C}$, the other one (figure 5) is.

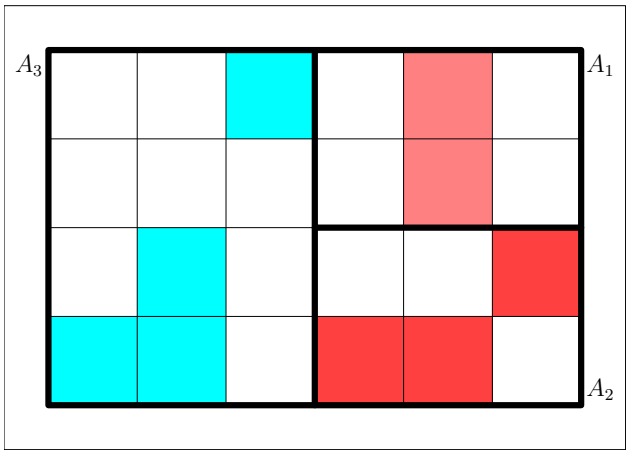

Figure 4: This choice of rectangles is not a member of $\mathcal{C}$ because the number of rectangles in A_2 is not a t-multiple of the number of rectangles in A_1, here with $s = 2, t = 3$ as in scenario III.

Let X be the random variable that corresponds to the ratio of the number of partition elements (rectangles) that are in A_3 and the total

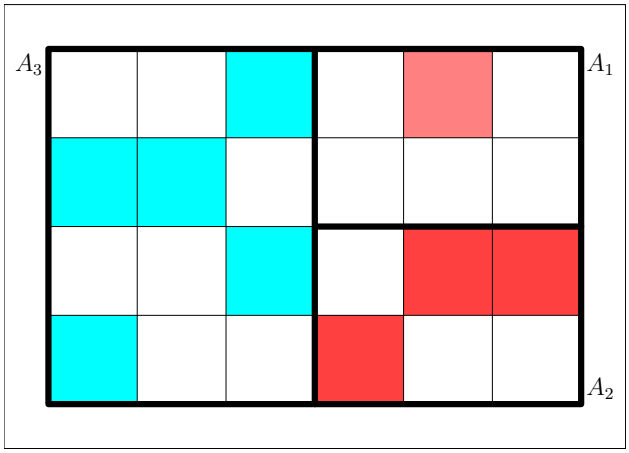

Figure 5: This choice of rectangles is a member of $\mathcal{C}$ because the number of rectangles in A_2 is a t-multiple of the number of rectangles in A_1, here with $s = 2, t = 3$ as in scenario III.

number of partition elements (rectangles) for a randomly chosen $C \in \mathcal{C}$. We would now anticipate that the expectation of X (which we will call EX) gives us an indication of the updated probability that Judy is in A_3 (so $EX \approx q_3$). The powerset approach is often superior to the uniformity approach (Grove and Halpern use uniformity, with all the necessary qualifications): if you have played Monopoly, you will know that the frequencies for rolling a 2, a 7, or a 10 with two dice tend to conform more closely to the binomial distribution (based on a powerset approach) rather than to the uniform distribution with $P(\text{rolling } i) = 1/11$ for $i = 2, \ldots, 12$.

Appendix II provides a formula for the powerset approach corresponding to the formula for the MAXENT approach, giving us q_3 dependent on t. Notice that this formula is for $t = 2, 3, 4, \ldots$. For $t = 1$ use the

Chu-Vandermonde identity to find that

$$EX_{12} = (t+1)\frac{\sum_{i=1}^{s} i\binom{ts}{i}\binom{ts}{ti}}{\sum_{i=0}^{s}\binom{ts}{i}\binom{ts}{ti}} = (t+1)\frac{s}{2}$$

and consequently $EX = 1/2$, as one would expect. For $t = 1/2, 1/3, 1/4, \ldots$ we can simply reverse the roles of A_1 and A_2. These results give us G_{pws} and a graph of the normalized odds vector (see figure 6), a bit bumpy around the middle because the t-values are discrete and farther apart in the middle, as $t = q/(1-q)$. Comparing the graphs of the normalized odds vector under Grove and Halpern's uniformity approach (G_{ind}), Jaynes' MAXENT approach (G_{max}), and the powerset approach suggested in this paper (G_{pws}), it is clear that the powerset approach supports MAXENT.

Going through the calculations, it seems at many places that the powerset approach should give its support to Grove and Halpern's uniformity approach in keeping with intuition T1. It was unexpected to find out that in the mathematical analysis $\alpha_{t,s}$ converges to a non-trivial factor and did not tend to negative or positive infinity, enabling a graph of the normalized odds vector that was not of the simple nature of the graph suggested by Grove and Halpern. Most surprisingly, the powerset approach, prima facie unrelated to an approach using information, supports the idea that a set of events about which nothing is known (such as A_3) gains in probability in the updated probability distribution compared to the set of events about which something is known (such as A_1 and A_2), even if it is

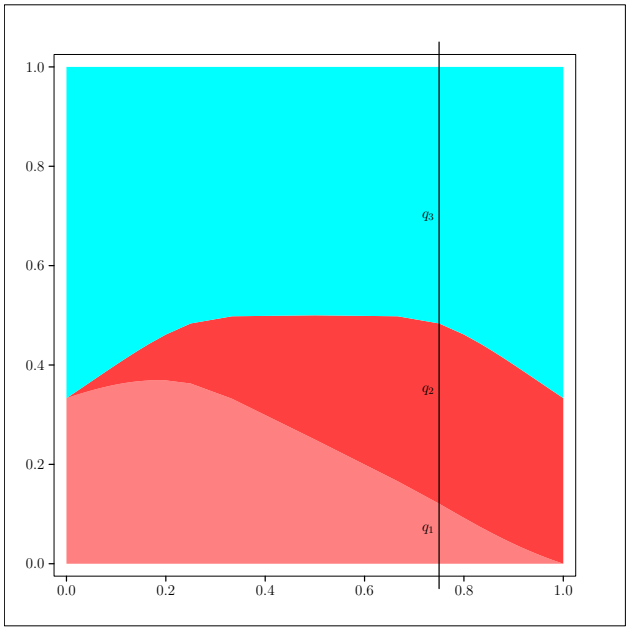

Figure 6: Judy Benjamin's updated probability assignment according to the powerset approach. $0 < q < 1$ forms the horizontal axis, the vertical axis shows the updated probability distribution (or the normalized odds vector) (q_1, q_2, q_3). The vertical line at $q = 0.75$ shows the specific updated probability distribution G_{pws} for the Judy Benjamin problem.

only partial information. Unless independence is specified, as in Sarah and sundowners at the Westcliff, the area of ignorance gains compared to the area of knowledge.

We now have several ways to characterize Judy's updated probabilities and updated probabilities following upon partial information in general. Only one of them, the uniformity approach, violates van Fraassen, Hughes,

and Harman's five symmetry requirements in (1986) and intuition T2. The uniformity approach, however, is the only one that satisfies intuition T1, an intuition which most people have when they first hear the story. Two arguments attenuate the position of the uniformity approach in comparison with the others.

First, T1 rests on an independence assumption which is not reflected in the problem. Although there is no indication that what Judy's commanders tell her is in any way dependent on her probability of being in Blue territory, it is not excluded either (see scenarios I–III earlier in this paper). MAXENT takes this uncertainty into consideration. Second, when we investigate the problem using the powerset approach it turns out that a division into equally probable, independent, and increasingly fine bits of information supports not intuition T1 but rather intuition T2. MAXENT, for now, is vindicated. We need to look for full employment not by looking at cleverly manipulating prior probabilities, but by making fresh observations, designing better experiments, and partitioning the theory space more finely.

5. Appendix I

Appendix I provides a concise but comprehensive summary of Jaynes' constraint rule not easily obtainable in the literature. Jaynes applied it to the Brandeis Dice Problem (see Jaynes, 1989, 243), but does not give a mathematical justification.

Let f be a probability distribution on a finite space $x_1, \ldots, x_m$ that fulfills the constraint

$$\sum_{i=1}^{m} r(x_i) f(x_i) = \alpha \tag{2}$$

An affine constraint can always be expressed by assigning a value to the expectation of a probability distribution (see Hobson, 1971). In Judy Benjamin's case, for example, let $r(x_1) = 0, r(x_2) = 1, r(x_3) = q$ and $\alpha = q$. Because f is a probability distribution it fulfills

$$\sum_{i=1}^{m} f(x_i) = 1 \tag{3}$$

We want to maximize Shannon's entropy, given the constraints (2) and (3),

$$-\sum_{i=1}^{m} f(x_i) \ln(x_i) \tag{4}$$

We use Lagrange multipliers to define the functional

$$J(f) = -\sum_{i=1}^{m} f(x_i) \ln f(x_i) + \lambda_0 \sum_{i=1}^{m} f(x_i) + \lambda_1 \sum_{i=1}^{m} r(x_i) f(x_i)$$

and differentiate it with respect to $f(x_i)$

$$\frac{\partial J}{\partial f(x_i)} = -\ln(f(x_i)) - 1 + \lambda_0 + \lambda_1 r(x_i) \tag{5}$$

Set (5) to 0 to find the necessary condition to maximize (4)

$$g(x_i) = e^{\lambda_0 - 1 + \lambda_1 r(x_i)}$$

This is the Gibbs distribution. We still need to do two things: (a) show that the entropy of g is maximal, and (b) show how to find λ_0 and λ_1. (a) is shown in Theorem 12.1.1 in Cover and Thomas (2006) and there is no reason to copy it here.

For (b), let

$$\lambda_1 = -\beta$$

$$Z(\beta) = \sum_{i=1}^{m} e^{-\beta r(x_i)}$$

$$\lambda_0 = 1 - \ln(Z(\beta))$$

To find λ_0 and λ_1 we introduce the constraint

$$-\frac{\partial}{\partial \beta} \ln(Z(\beta)) = \alpha$$

To see how this constraint gives us λ_0 and λ_1, Jaynes' solution of the Brandeis Dice Problem (see Jaynes, 1989, 243) is a helpful example. We are, however, interested in a general proof that this choice of λ_0 and λ_1 gives us

the probability distribution maximizing the entropy. That g so defined maximizes the entropy is shown in (a). We need to make sure, however, that with this choice of λ_0 and λ_1 the constraints (2) and (3) are also fulfilled.

First, we show

$$\sum_{i=1}^{m} g(x_i) = \sum_{i=1}^{m} e^{\lambda_0 - 1 + \lambda_1 r(x_i)} = e^{\lambda_0 - 1} \sum_{i=1}^{m} e^{\lambda_1 r(x_i)} =$$

$$e^{-\ln(Z(\beta))} Z(\beta) = 1$$

Then, we show, by differentiating $\ln(Z(\beta))$ using the substitution $x = e^{-\beta}$

$$\alpha = -\frac{\partial}{\partial \beta} \ln(Z(\beta)) = -\frac{1}{\sum_{i=1}^{m} x^{r(x_i)}} \left(\sum_{i=1}^{m} r(x_i) x^{r(x_i)-1} \right) (-x) =$$

$$\frac{\sum_{i=1}^{m} r(x_i) x^{r(x_i)}}{\sum_{i=1}^{m} x^{r(x_i)}}$$

And, finally,

$$\sum_{i=1}^{m} r(x_i) g(x_i) = \sum_{i=1}^{m} r(x_i) e^{\lambda_0 - 1 + \lambda_1 r(x_1)} = e^{\lambda_0 - 1} \sum_{i=1}^{m} r(x_i) e^{\lambda_1 r(x_1)} =$$

$$e^{\lambda_0 - 1} \sum_{i=1}^{m} r(x_i) x^{r(x_i)} = \alpha e^{\lambda_0 - 1} \sum_{i=1}^{m} x^{r(x_i)} = \alpha e^{\lambda_0 - 1} \sum_{i=1}^{m} e^{-\beta r(x_i)} =$$

$$\alpha Z(\beta) e^{\lambda_0 - 1} = \alpha Z(\beta)) e^{-\ln(Z(\beta))} = \alpha$$

Filling in the variables from Judy Benjamin's scenario gives us result (1).

The lambdas are:

$$\lambda_0 = 1 - \ln\left(\sum_{i=1}^{m} e^{\lambda_1 r(x_i)}\right) \qquad \lambda_1 = \ln q - \ln(1-q)$$

We combine the normalized odds vector $(0.16, 0.48, 0.36)$ following from these lambdas using Dempster's Rule of Combination with (MAP) and get result (1).

6. Appendix II

The binomial distribution dictates the value of EX, using simple combinatorics. In this case we require, again for convenience, that n be divisible by t and the 'grain' of the partition A be $s = n/t$. We introduce a few variables which later on will help for abbreviation:

$$n = ts \qquad 2m = n \qquad 2j = n - 1 \qquad T = t^2 + 1$$

EX, of course, depends both on the grain of A and the value of t. It makes sense to make it independent of the grain by letting the grain become increasingly finer and by determining EX as $s \to \infty$. This cannot be done for the binomial distribution, as it is notoriously uncomputable for large numbers (even with a powerful computer things get dicey around $s = 10$). But, equally notorious, the normal distribution provides a good approximation of the binomial distribution and will help us arrive at a formula for G_{pws} (corresponding to G_{ind} and G_{max}), determining the value q_3

dependent on q as suggested by the powerset approach.

First, we express the random variable X by the two independent random variables X_{12} and X_3. X_{12} is the number of partition elements in the randomly chosen C which are either in A_1 or in A_2 (the random variable of the number of partition elements in A_1 and the random variable of the number of partition elements in A_2 are decisively not independent, because they need to obey (HDQ)); X_3 is the number of partition elements in the randomly chosen C which are in A_3. A relatively simple calculation shows that $EX_3 = n$, which is just what we would expect (either the powerset approach or the uniformity approach would give us this result):

$$EX_3 = 2^{-2n} \sum_{i=0}^{2n} i \binom{2n}{i} = n \text{ (use } \binom{n}{k} = \frac{n}{k}\binom{n-1}{k-1}\text{)}$$

The expectation of X, X being the random variable expressing the ratio of the number of sets covering A_3 and the number of sets covering $A_1 \cup A_2 \cup A_3$, is

$$EX = \frac{EX_3}{EX_{12} + EX_3} = \frac{n}{EX_{12} + n}$$

If we were able to use uniformity and independence, $EX_{12} = n$ and $EX = 1/2$, just as Grove and Halpern suggest (although their uniformity approach is admittedly less crude than the one used here). Will the powerset approach concur with the uniformity approach, will it support the principle of maximum entropy, or will it make another suggestion on how to update the prior probabilities? To answer this question, we must find out

what EX_{12} is, for a given value t and $s \to \infty$, using the binomial distribution and its approximation by the normal distribution.

Using combinatorics,

$$EX_{12} = (t+1)\frac{\sum_{i=1}^{s} i\binom{ts}{i}\binom{ts}{ti}}{\sum_{i=0}^{s}\binom{ts}{i}\binom{ts}{ti}}$$

Let us call the numerator of this fraction NUM and the denominator DEN. According to the de Moivre-Laplace Theorem,

$$\text{DEN} = \sum_{i=0}^{s}\binom{ts}{i}\binom{ts}{ti} \approx 2^{2n}\sum_{i=0}^{s}\int_{i-\frac{1}{2}}^{i+\frac{1}{2}} \mathcal{N}(\frac{n}{2},\frac{n}{4})(i)\mathcal{N}(\frac{n}{2},\frac{n}{4})(ti)di$$

where

$$\mathcal{N}(\mu,\sigma^2)(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Substitution yields

$$\text{DEN} \approx 2^{2n}\frac{1}{\pi m}\sum_{i=0}^{s}\int_{i-\frac{1}{2}}^{i+\frac{1}{2}}\exp\left(-\frac{(x-m)^2}{m} - \frac{t^2\left(x-\frac{m}{t}\right)^2}{m}\right)dx$$

Consider briefly the argument of the exponential function:

$$-\frac{(x-m)^2}{m} - \frac{t^2\left(x-\frac{m}{t}\right)^2}{m} = -\frac{t^2}{m}(a''x^2+b''x+c'') = -\frac{t^2}{m}\left(a''(x-h'')^2+k''\right)$$

with (the double prime sign corresponds to the simple prime sign for the

numerator later on)

$$a'' = \frac{1}{t^2}T \qquad b'' = (-2m)\frac{1}{t^2}(t+1) \qquad c'' = 2m^2\frac{1}{t^2}$$

$$h'' = -b''/2a'' \qquad k'' = a''h''^2 + b''h'' + c''$$

Consequently,

$$\text{DEN} \approx 2^{2n}\exp\left(-\frac{t^2}{m}k''\right)\sqrt{\frac{1}{\pi a'' m t^2}}\int_{-\infty}^{s+\frac{1}{2}}\mathcal{N}\left(h'', \frac{m}{2a''t^2}\right)dx$$

And, using the error function for the cumulative density function of the normal distribution,

$$\text{DEN} \approx 2^{2n-1}\sqrt{\frac{1}{\pi a'' m t^2}}\exp\left(-\frac{k''t^2}{m}\right)(1-\text{erf}(w'')) \tag{6}$$

with

$$w'' = \frac{t\sqrt{a''}\left(s+\frac{1}{2}-h''\right)}{\sqrt{m}}$$

We proceed likewise with the numerator, although the additional factor introduces a small complication:

$$\begin{aligned}\text{NUM} &= \sum_{i=1}^{s} i\binom{ts}{i}\binom{ts}{ti} = \sum_{i=1}^{s} s\binom{ts}{i}\binom{ts-1}{ti-1} \\ &\approx s2^{2n-1}\sum_{i=1}^{s}\mathcal{N}\left(m, \frac{m}{2}\right)(i)\mathcal{N}\left(j, \frac{j}{2}\right)(ti-1)\end{aligned}$$

Again, we substitute and get

$$\text{NUM} \approx s2^{2n-1}\left(\pi\sqrt{mj}\right)^{-1}\sum_{0}^{s-1}\int_{i-\frac{1}{2}}^{i+\frac{1}{2}}\exp\left(a'(x-h')^2+k'\right)$$

where the argument for the exponential function is

$$-\frac{1}{mj}\left(j(x-m)^2+mt^2\left(x-\frac{j+1}{t}\right)^2\right)$$

and therefore

$$a'=j+mt^2 \qquad b'=2j(1-m)+2mt\,(t-j) \qquad c'=j(1-m)^2+m\,(t-j-1)^2$$

$$h'=-b'/2a' \qquad k'=a'h'^2+b'h'+c'$$

Using the error function,

$$\text{NUM} \approx 2^{2n-2}\frac{s}{\sqrt{\pi a'}}\exp\left(-\frac{k'}{mj}\right)(1+\text{erf}(w')) \tag{7}$$

with

$$w'=\frac{\sqrt{a'}\left(s-\frac{1}{2}-h'\right)}{\sqrt{mj}}$$

Combining (6) and (7),

$$\begin{aligned} EX_{12} &= (t+1)\frac{\text{NUM}}{\text{DEN}} \\ &\approx \frac{1}{2}(t+1)\sqrt{\frac{Tts}{Tts-1}}se^{\alpha_{t,s}} \end{aligned}$$

for large s, because the arguments for the error function w' and w'' escape to positive infinity in both cases (NUM and DEN) so that their ratio goes to 1. The argument for the exponential function is

$$\alpha_{t,s} = -\frac{k'}{mj} + \frac{k''t^2}{m}$$

and, for $s \to \infty$, goes to

$$\alpha_t = \frac{1}{2}T^{-2}(2t^3 - 3t^2 + 4t - 5)$$

Notice that, for $t \to \infty$, α_t goes to 0 and

$$EX = \frac{n}{EX_{12} + n} \to \frac{2}{3}$$

in accordance with intuition T2.

Bradley, R., 2005. Radical Probabilism and Bayesian Conditioning. *Philosophy of Science*, 72(2):342–364.

Cover, T. and Thomas, J., 2006. *Elements of Information Theory*, volume 6. Wiley, Hoboken, NJ.

Csiszár, I., 1967. Information-Type Measures of Difference of Probability Distributions and Indirect Observations. *Studia Scientiarum Mathematicarum Hungarica*, 2:299–318.

Diaconis, P. and Zabell, S., 1982. Updating Subjective Probability. *Journal of the American Statistical Association*, pages 822–830.

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

Douven, I. and Romeijn, J., 2009. A New Resolution of the Judy Benjamin Problem. *CPNSS Working Paper*, 5(7):1–22.

Gardner, M., 1959. Mathematical Games. *Scientific American*, 201(4):174–182.

Grove, A. and Halpern, J., 1997. Probability Update: Conditioning Vs. Cross-Entropy. In *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence*. Citeseer, Providence, Rhode Island.

Grünwald, P. and Halpern, J., 2003. Updating Probabilities. *Journal of Artificial Intelligence Research*, 19:243–278.

Halpern, J. Y., 2003. *Reasoning About Uncertainty*. MIT Press, Cambridge, MA.

Hobson, A., 1971. *Concepts in Statistical Mechanics*. Gordon and Beach, New York, NY.

Jaynes, E., 1989. *Papers on Probability, Statistics and Statistical Physics*. Springer, Dordrecht.

Jaynes, E. and Bretthorst, G., 1998. *Probability Theory: the Logic of Science*. Cambridge University Press, Cambridge, UK.

Jeffrey, R., 1965. *The Logic of Decision*. McGraw-Hill, New York, NY.

Seidenfeld, T., 1979. Why I Am Not an Objective Bayesian; Some Reflections Prompted by Rosenkrantz. *Theory and Decision*, 11(4):413–440.

Uffink, J., 1995. Can the Maximum Entropy Principle be Explained as a Consistency Requirement? *Studies In History and Philosophy of Science Part B: Studies In History and Philosophy of Modern Physics*, 26(3):223–261.

Uffink, J., 1996. The Constraint Rule of the Maximum Entropy Principle. *Studies In History and Philosophy of Science Part B: Studies In History and Philosophy of Modern Physics*, 27(1):47–79.

Van Fraassen, B., 1981. A Problem for Relative Information Minimizers in Probability Kinematics. *The British Journal for the Philosophy of Science*, 32(4):375–379.

Van Fraassen, B.; Hughes, R.; and Harman, G., 1986. A Problem for Relative Information Minimizers, Continued. *The British Journal for the Philosophy of Science*, 37(4):453–463.