Synthese

The Principle of Maximum Entropy and a Problem in Probability Kinematics
--Manuscript Draft--

Manuscript Number:	
Full Title:	The Principle of Maximum Entropy and a Problem in Probability Kinematics
Article Type:	Original Research
Keywords:	Judy Benjamin; Principle of Maximum Entropy; Coarsening at Random; Full Employment Theorem; Probability Kinematics; Epistemic Entrenchment
Corresponding Author:	Stefan Lukits University of British Columbia Vancouver, British Columbia CANADA
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	University of British Columbia
Corresponding Author's Secondary Institution:	
First Author:	Stefan Lukits
First Author Secondary Information:	
Order of Authors:	Stefan Lukits
Order of Authors Secondary Information:	
Abstract:	Given a set of probabilities on an event space and a new observation, it is common to form updated probabilities by using rules of conditioning. Some observations pose constraints that cannot be addressed by standard conditional probabilities or Jeffrey conditioning. The principle of maximum entropy claims that it is then appropriate to form an updated probability assessment by minimizing information gain consistent with the observation. Opposition to the principle of maximum entropy leans heavily on a counterexample: the Judy Benjamin problem. The article shows that an intuitively plausible approach, which on account of independence assumptions should support the opponents, turns out to support the principle of maximum entropy. It also demonstrates that the opponents improperly apply independence assumptions to the problem.

"The Principle of Maximum Entropy and a Problem in Probability Kinematics"

Stefan Lukits

University of British Columbia

20 22nd Avenue East

Vancouver BC V5V 1T4

saiserit@streetgreek.com

(604) 321-3440

The Principle of Maximum Entropy and a Problem in Probability Kinematics

Stefan Lukits

Abstract Given a set of probabilities on an event space and a new observation, it is common to form updated probabilities by using rules of conditioning. Some observations pose constraints that cannot be addressed by standard conditional probabilities or Jeffrey conditioning. The principle of maximum entropy claims that it is then appropriate to form an updated probability assessment by minimizing information gain consistent with the observation. Opposition to the principle of maximum entropy leans heavily on a counterexample: the Judy Benjamin problem. The article shows that an intuitively plausible approach, which on account of independence assumptions should support the opponents, turns out to support the principle of maximum entropy. It also demonstrates that the opponents improperly apply independence assumptions to the problem.

Keywords Judy Benjamin · Principle of Maximum Entropy · Coarsening at Random · Full Employment Theorem · Probability Kinematics · Epistemic Entrenchment

1 Introduction

Probability kinematics is the field of inquiry asking how we should update a probability distribution in the light of evidence. If the evidence comes as an event, it is relatively uncontroversial to use conditional probabilities (call this standard conditioning). Sometimes, however, the evidence may not relate the certainty of an event but a reassessment of its uncertainty or its probabilistic

Stefan Lukits
University of British Columbia
Department of Philosophy
1866 Main Mall E370
Vancouver BC Canada V6T 1Z1
E-mail: saiserit@streetgreek.com

relation to other events (see [Jeffrey(1965), 153ff]), expressible in a shift in expectation (see [Hobson(1971)]). Bas van Fraassen has come up with an 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.
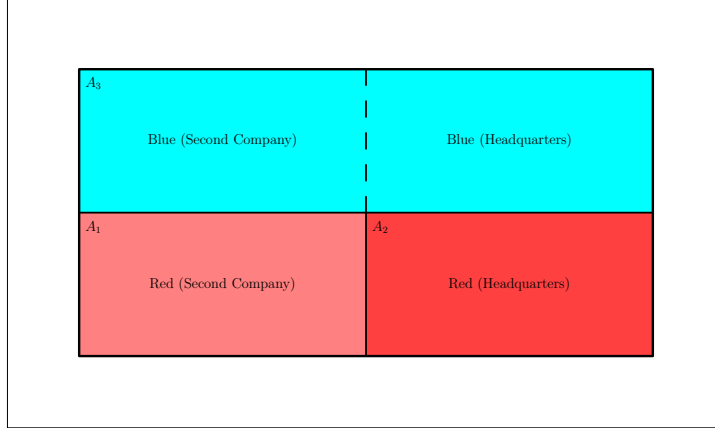

Fig. 1 Judy Benjamin's map. Blue territory (A_3) is friendly and does not need to be divided into a Headquarters and a Second Company area.

In van Fraassen's interpretation of 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 of which we are certain. Grove and Halpern have written an article [Grove and Halpern(1997)] 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.

If we assume that the attempt fails to define an event on which Judy Benjamin could condition her probabilities, we are left with two possibilities. Her new information (it is three times as likely to land in A_2 than to land in A_1,

see figure 1 and the details of the problem in the next section) may mean that we have a redistribution of a complete partition of the probabilities. This is called Jeffrey conditioning and calls for Jeffrey's rule. Jeffrey's rule is contested in some circles, but we will for this project accept its validity in probability kinematics. We will see in what follows that some make the case that Jeffrey conditioning is the correct way to solve the Judy Benjamin problem.

The third possibility to solve this problem (after standard conditioning and Jeffrey's rule) is to consult a highly contested procedure to find an objective updating procedure in case the first two scenarios do not apply: the principle of maximum entropy (from now on MAXENT for abbreviation). MAXENT can be applied to any situation in which we have a completely quantified probability distribution and an affine constraint (we will explain the nature of affine constraints in more detail later). If our new evidence is the observation of an event (or simply certainty about an event that we did not have previously), then the event provides an affine constraint and can be used for updating by means of standard conditioning. If our new evidence is a redistribution of probabilities where we can apply Jeffrey's rule, then the redistribution provides an affine constraint and can be used for updating by means of Jeffrey's rule. These two possibilities, however, do not exhaust affine constraints. The Judy Benjamin problem illustrates the third possibility where the affine constraint only redistributes some groups of probabilities and leaves open the question how this will affect the probabilities not included in this redistribution.

Advocates of MAXENT claim that in this case the probabilities should be adjusted so that they are minimally affected (we make this precise by using information theory) while at the same time according with the constraint. Opponents of this view grant that MAXENT is an important tool of probability kinematics, but that results of MAXENT that are difficult to accept (such as in the Judy Benjamin case) urge us to embrace a more pluralistic, situation-specific methodology.

Joseph Halpern, for example, 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 MAXENT 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" [Howson and Franklin(1994), 456] is also what Colin Howson and Allan Franklin advise about the more basic 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 ... may be an art, or a skill, rather than a mat-

ter of calculation or derivation from more fundamental epistemic principles" [Douven and Romeijn(2009), 16] (for the Bradley quote see [Bradley(2005), 362]).

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.

We have no proof that MAXENT is the only rationally defensible objective method to update probabilities given an affine constraint. The literature outlines many of the 'nice properties' (so-called by an anonymous reviewer) of MAXENT: It seamlessly generalizes standard conditioning and Jeffrey's rule where they are applicable (see [Caticha and Giffin(2006)]). It underlies the entropy concentration phenomenon described in Jaynes' standard work *Probability Theory: the Logic of Science*, which contains other arguments in favour of MAXENT, some of which you may recognize by family resemblance in the rest of this paper. Entropy concentration refers to the unique property of the MAXENT solution to have other distributions which obey the affine constraint cluster around it. Shore and Johnston have shown that under certain rationality assumptions MAXENT is the unique solution problems of probability updating (see [Shore and Johnson(1980)]). When used to make predictions whose quality is measured by a logarithmic score functions, posterior probabilities provided by MAXENT result in minimax optimal decisions (see [Topsøe(1979)], [Walley(1991)], and [Grünwald(2000)]). Under a logarithmic scoring rule these posterior probabilities are in some sense optimal.

Despite all these nice properties, we want the reader to follow us in a more simple line of argument. When new evidence is provided to us, it is rational to adjust our beliefs minimally in light of it. We do not want to draw out more information from the new evidence than necessary. We are the first to

admit that there are numerous problems here that need addressing. What do we mean by rationality? What are the semantics of the word 'minimal'? What are the formal properties of such posterior probabilities? Are they unique? Are they compatible with other intuitive methods of updating? Are there counter-intuitive examples that would encourage us to give up on this line of thought rather than live with its consequences? Given some decent answers to these questions, however, we feel that MAXENT cuts a good figure as a first pass to provide objective solutions to these types of problems, and the burden on opponents who usually deny that there are such objective solutions exist grows heavy.

One particular emphasis in this paper is that the reasoning of the opponents of MAXENT in the Judy Benjamin case is flawed because they make independence assumptions that on closer inspection do not hold up. We provide a number of scenarios which easily pass the information given in the problem which violate these independence assumptions. That does not mean that the information given in the problem suggests these scenarios, it only means that we are not entitled to make independence assumptions. That, in turn, does not privilege the MAXENT solution, although MAXENT does not lean on independence assumptions that other solutions illegitimately make. For MAXENT confronts us with a much stronger claim than merely providing a passable or useful solution to the Judy Benjamin problem: it lays claim to much greater generality and, abhorred by many formal epistemologists, to objectivity. These claims must be motivated elsewhere—we are only showing that opponents cannot claim an easy victory by pulling out old Judy Benjamin, although this is still widely practised.

Dealing in stereotypes for a moment, there is a long-standing disagreement between philosophers on the one hand and physicists on the other hand whether (the philosophers) updating probabilities is irreducibly accompanied by thoughtful deliberation choosing between approaches depending on individual problems, or (the physicists) problems are ill-posed if they do not contain the information necessary to let a non-arbitrary, objective process arrive at a unique updated probability distribution. 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 proven 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),

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

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

(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 redistribution 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 [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 observation 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. (These considerations are strong enough that Luc Bovens uses them as an assumption to solve Adam Elga's Sleeping Beauty problem by parity of reasoning in [Bovens(2010)].)

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 [Grove and Halpern(1997)] as well as Douven and Romeijn [Douven and Romeijn(2009)] make a case for this distribution. Grove and Halpern use standard conditioning on the event of the message being transmitted to Judy. Douven and Romeijn use 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. [Van Fraassen(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 the unique updated probability assignment according with these assumptions. 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. 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 constraint be affine, see [Csiszár(1967)]).

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

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

The updated probability of being on Blue territory (A_3) has increased from 50% to approximately 53%. Grove and Halpern find this result "highly counter-intuitive" [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 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.). Alternatively, if circumstances are favourable (as they are in Judy Benjamin's case), we may use the Kullback-Leibler Divergence and differentiate it to obtain where it is minimal.

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 is dropped off by a pilot who flips two coins. If the first coin lands H, then Judy is dropped off in Blue territory, otherwise in Red territory. If the second coin lands H, she is dropped off in the Headquarters area, otherwise in the Second Company area. Judy's commanders find out that the second coin is biased $q : 1-q$ toward H with $q = 0.75$. The normalized odds vector is $G_I(0.75) = (0.125, 0.375, 0.500)$ and agrees with T1, because
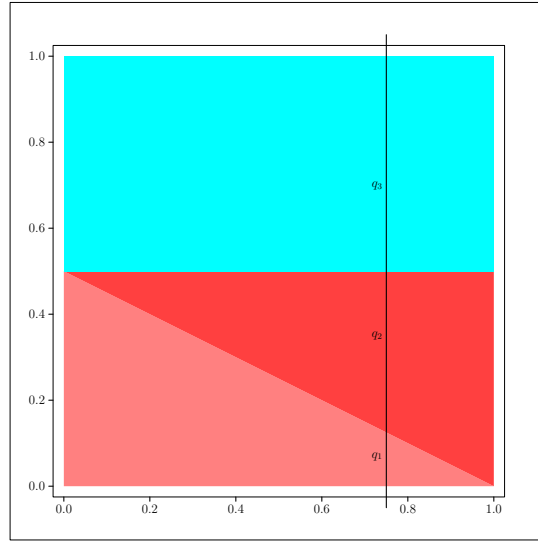
Fig. 2 Judy Benjamin's updated probability assignment according to intuition T1. $0 < q < 1$ forms the horizontal axis, the vertical axis shows the updated probability distribution (or the normalized odds vector) (q_1, q_2, q_3). The vertical line at $q = 0.75$ shows the specific updated probability distribution $G_{\mathrm{ind}}(0.75)$ for the Judy Benjamin problem.

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 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 $G_{\mathrm{II}}(0.75) = (0.1, 0.3, 0.6)$ 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 $G_{\mathrm{III}}(0.75) \approx (.108, .324, .568)$ (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),
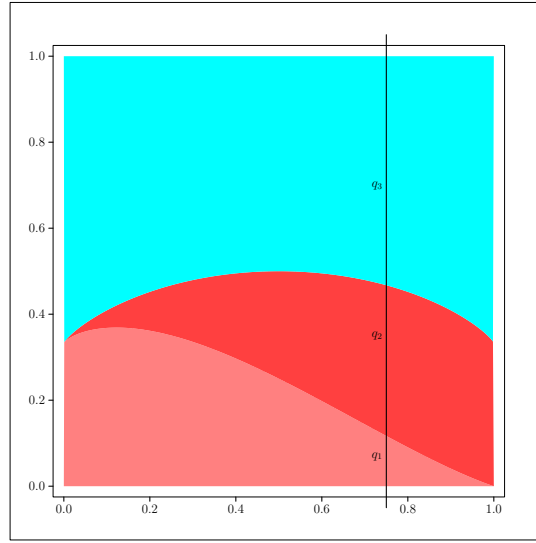
Fig. 3 Judy Benjamin's updated probability assignment using MAXENT. $0 < q < 1$ forms the horizontal axis, the vertical axis shows the updated probability distribution (or the normalized odds vector) (q_1, q_2, q_3). The vertical line at $q = 0.75$ shows the specific updated probability distribution $G_{\max}(0.75)$ for the Judy Benjamin problem.

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 makes clear. This is not to say that scenario II is the scenario that pertains in Judy Benjamin's case. It only says that there is no natural assumption in Judy Benjamin's case that the probabilities are independent of each other in light of the new evidence, for scenario II is perfectly natural (whether it is true or not is a completely different question) and reveals how dependence can naturally flow from the information that Judy Benjamin receives.

Douven and Romeijn's strong independence claim relying on intuition T1 leads them to apply Jeffrey's rule to the Judy Benjamin problem with the additional constraint $q_3 = p_3$. 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]). The idea of epistemic entrenchment is at

Epistemic entrenchment	q_1	q_2	q_3
with respect to A_1	1/4	3/4	0
with respect to A_2	1/12	1/4	2/3
with respect to A_3	1/8	3/8	1/2

odds with MAXENT and seems to imply just what the full employment theorem claims: judgments so framed "will depend on the judgmental skills of the agent, typically acquired not in the inductive logic class but by subject specific training" [Bradley(2005), 349]. To pursue the relation between the semantics of conditionals, MAXENT, and the full employment theorem would take us too far afield at present and shall be undertaken elsewhere. For the Judy Benjamin problem, it is not clear why Douven and Romeijn think that the way the problem is posed implies a strong epistemic entrenchment for Adams conditioning (Adams conditioning is the kind of conditioning that will leave the antecedent alone). Scenarios II-III provide realistic alternatives where Adams conditioning is inappropriate.

Judy Benjamin may also receive (HDQ) because her informers have found out that Red Headquarters troops have occupied the entire Blue territory ($q_1 = 3p_1, q_2 = p_2, q_3 = 0$, the epistemic entrenchment is with respect to q_2); because they have found out that Blue troops have occupied two-thirds of the Red Second Company area ($q_1 = p_1, q_2 = (1/3)p_2, q_3 = (4/3)p_3$, the epistemic entrenchment is with respect to q_1); or because they have found out that Red Headquarters troops have taken over half of the Red Second Company area ($q_1 = (1/2)p_1, q_2 = (3/2)p_2, q_3 = p_3$, the epistemic entrenchment is with respect to q_3 and what Douven and Romeijn take to be an assumption in the wording of the problem). There is nothing in the problem that supports Douven and Romeijn's narrowing of the options. The table reiterates these options, with the third, shaded line representing intuition (T1) and the epistemic entrenchment defended by Douven and Romeijn.

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, standard conditioning) MAX-ENT "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 counter-intuitive 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 of the Blue and Red territories into sets of equal measure (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 $A_1 \cup A_2$. Initially, any subset of this partition is a candidate for Judy Benjamin to consider. The constraint imposed by (HDQ) is that now we only consider subsets for which there are three times as many partition sets (or rectangles, although we are not necessarily limiting ourselves to rectangles) in A_2 than there are in A_1. In figures 4 and 5 there are diagrams of two subsets. One of them (figure 4) is not a candidate, the other one (figure 5) is.
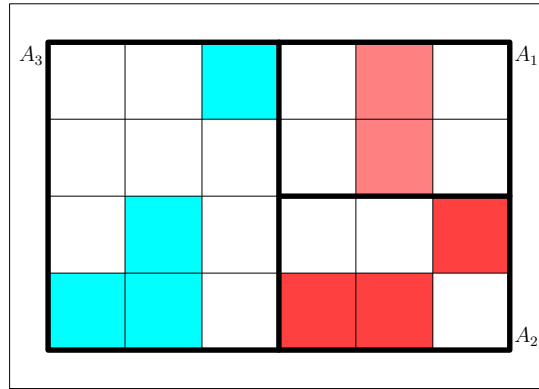

Fig. 4 This choice of rectangles is not a candidate because the number of rectangles in A_2 is not a t-multiple of the number of rectangles in A_1, here with $s = 2, t = 3$ as in scenario III.

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 number of partition elements (rectangles) for a randomly chosen candidate. 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 $i = 2, \ldots, 12$.

Calculations not provided here 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
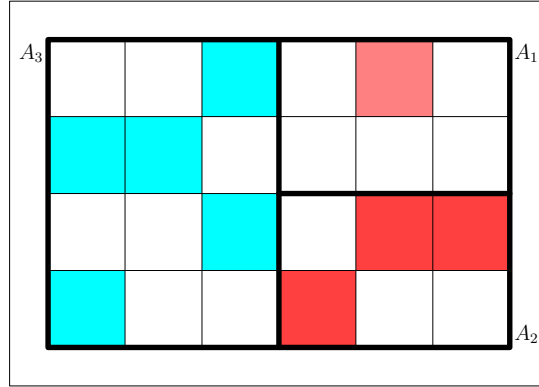
Fig. 5 This choice of rectangles is a candidate because the number of rectangles in A_2 is a t-multiple of the number of rectangles in A_1, here with $s = 2, t = 3$ as in scenario III.

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 is unexpected to find out that in the mathematical analysis the parameters converge to a non-trivial factor and do not tend to negative or positive infinity, enabling a graph of the normalized odds vector that is 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 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 [Van Fraassen(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
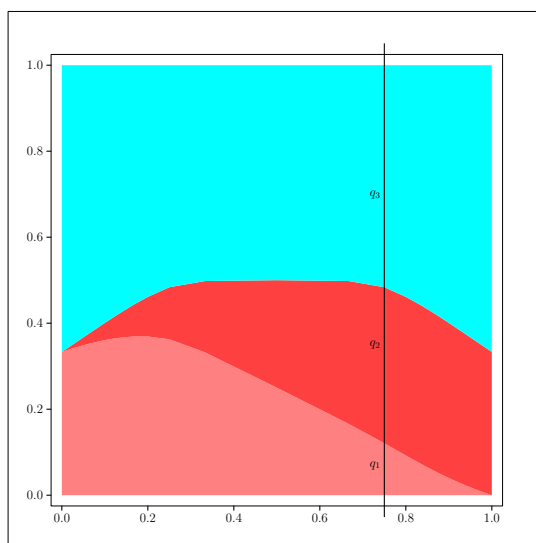
Fig. 6 Judy Benjamin's updated probability assignment according to the powerset approach. $0 < q < 1$ forms the horizontal axis, the vertical axis shows the updated probability distribution (or the normalized odds vector) (q_1, q_2, q_3). The vertical line at $q = 0.75$ shows the specific updated probability distribution G_{pws} for the Judy Benjamin problem.

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 cleverly manipulating prior probabilities, but by making fresh observations, designing better experiments, and partitioning the theory space more finely.

References

[Bovens(2010)] Bovens L (2010) Judy benjamin is a sleeping beauty. Analysis 70(1):23–26

[Bradley(2005)] Bradley R (2005) Radical probabilism and bayesian conditioning. Philosophy of Science 72(2):342–364

[Caticha and Giffin(2006)] Caticha A, Giffin A (2006) Updating probabilities. In: MaxEnt 2006, the 26th International Workshop on Bayesian Inference and Maximum Entropy Methods

[Cover and Thomas(2006)] Cover T, Thomas J (2006) Elements of Information Theory, vol 6. Wiley, Hoboken, NJ

[Csiszár(1967)] Csiszár I (1967) Information-type measures of difference of probability distributions and indirect observations. Studia Scientiarum Mathematicarum Hungarica 2:299–318

[Diaconis and Zabell(1982)] Diaconis P, Zabell S (1982) Updating subjective probability. Journal of the American Statistical Association pp 822–830

[Douven and Romeijn(2009)] Douven I, Romeijn J (2009) A new resolution of the judy benjamin problem. CPNSS Working Paper 5(7):1–22

[Gardner(1959)] Gardner M (1959) Mathematical games. Scientific American 201(4):174–182

[Grove and Halpern(1997)] Grove A, 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(2000)] Grünwald P (2000) Maximum entropy and the glasses you are looking through. In: Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence, Morgan Kaufmann Publishers, pp 238–246

[Grünwald and Halpern(2003)] Grünwald P, Halpern J (2003) Updating probabilities. Journal of Artificial Intelligence Research 19:243–278

[Halpern(2003)] Halpern JY (2003) Reasoning About Uncertainty. MIT Press, Cambridge, MA

[Hobson(1971)] Hobson A (1971) Concepts in Statistical Mechanics. Gordon and Beach, New York, NY

[Howson and Franklin(1994)] Howson C, Franklin A (1994) Bayesian conditionalization and probability kinematics. The British Journal for the Philosophy of Science 45(2):451–466

[Jaynes and Bretthorst(1998)] Jaynes E, Bretthorst G (1998) Probability Theory: the Logic of Science. Cambridge University Press, Cambridge, UK

[Jeffrey(1965)] Jeffrey R (1965) The Logic of Decision. McGraw-Hill, New York, NY

[Shore and Johnson(1980)] Shore J, Johnson R (1980) Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. IEEE Transactions on Information Theory 26(1):26–37

[Topsøe(1979)] Topsøe F (1979) Information-theoretical optimization techniques. Kybernetika 15(1):8–27

[Uffink(1996)] 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(1981)] 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(1986)] Van Fraassen B (1986) A problem for relative information minimizers, continued. The British Journal for the Philosophy of Science 37(4):453–463

[Walley(1991)] Walley P (1991) Statistical Reasoning with Imprecise Probabilities. Chapman and Hall London