Judy Benjamin and Halpern's Full Employment Theory

Annual Meeting of the Canadian Society for the History and Philosophy of Science

Stefan Lukits

May 29, 2011

Halpern's Full Employment Theorem

Choosing a representation:

- possible worlds
- probability measures
- lower and upper probabilities
- Dempster-Shafer belief functions
- possibility measures
- ranking functions
- relative likelihoods
- plausibility measures

Grove and Halpern

"one must always think carefully about precisely what the information means"
("Probability Update . . . ," p6)

Joseph Halpern

"there is no escaping the need to understand the details of the application"
(*Reasoning . . .*, p423)

Pluralism Versus Statistical Physics I

Statistical Physics

"Jaynes's principle of maximum entropy and Kullbacks principle of minimum cross-entropy (minimum directed divergence) are shown to be uniquely correct methods for inductive inference when new information is given in the form of expected values." (Shore and Johnson)

Pluralism

"The uniqueness proofs are flawed, or rest on unreasonably strong assumptions. A more general class of inference rules, maximizing the so-called Rényi entropies, is exhibited which also fulfill the reasonable part of the consistency assumptions." (Jos Uffink)

Pluralism Versus Statistical Physics II

Statistical Physics

"We show that Skilling's method of induction leads us to a unique general theory of inductive inference, the maximum entropy method, and precisely how it is that other entropies such as those of Rényi or Tsallis are ruled out for problems of inference. We then explore the compatibility of Bayes and maximum entropy updating. We show that maximum entropy is capable of producing every aspect of orthodox Bayesian inference and prove the complete compatibility of Bayesian and entropy methods." (Adom Giffin)

Pluralism Versus Statistical Physics III

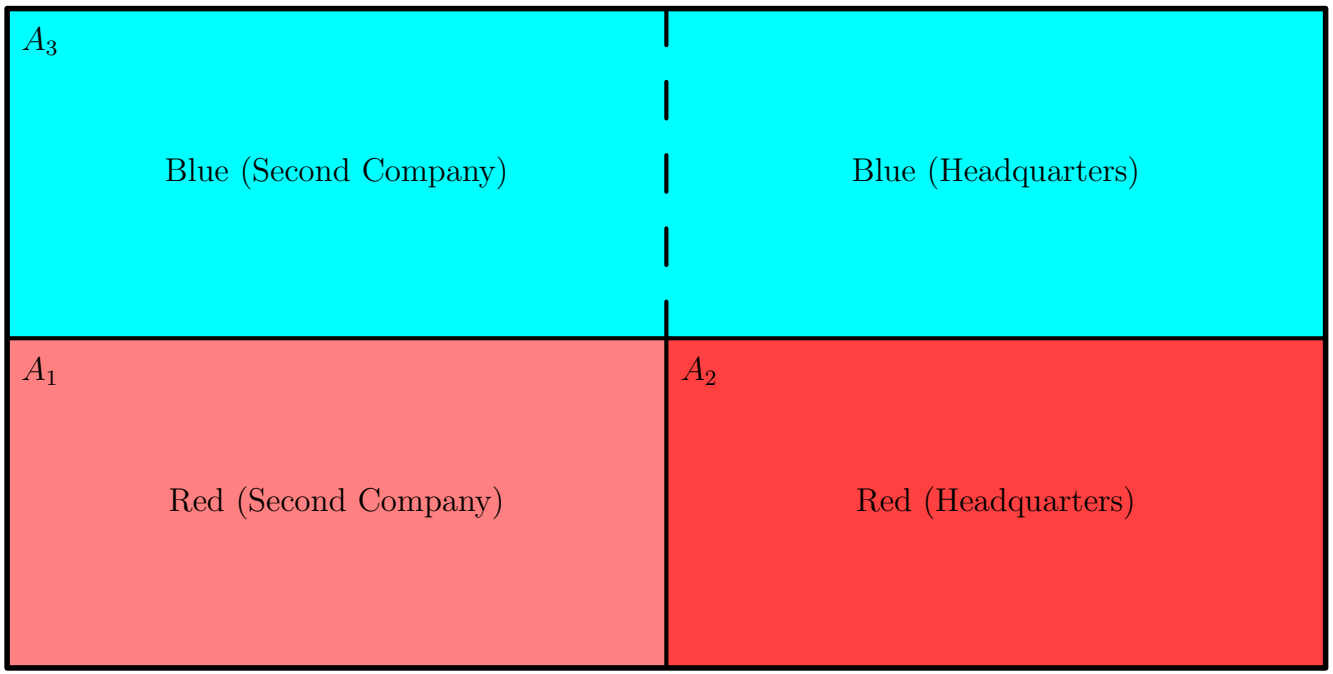
Pluralism

"It is perhaps not surprising that there are proponents of maximum entropy and relative entropy who recommend that if an agent's information can be characterized by a set C of constraints, then the agent should act 'as if' the probability is determined by the measure that maximizes entropy relative to C (i.e., the measure that has the highest entropy of all the measures in C). Similarly, if the agent starts with a particular measure μ and gets new information characterized by C, he should update to the measure μ' that satisfies C such that the relative entropy between μ' and μ is a minimum. Maximum entropy and relative entropy have proved quite successful in a number of applications, from physics to natural-language modeling. Unfortunately, they also exhibit some counterintuitive behavior on certain applications. Although they are valuable tools, they should be used with care."
(Joseph Halpern)

Objective

- → Halpern claims that the principle of maximum entropy sometimes exhibits counterintuitive behaviour on certain applications. Therefore, the principle of full employment applies.
- → Halpern singles out the Judy Benjamin case to demonstrate the counterintuitive behaviour of maximum entropy.
- → We will show that the Judy Benjamin case does not behave counterintuitively when approached via maximum entropy. On the contrary, maximum entropy reveals where the information given by the case is incompletely specified and provides precisely the results that correspond to our intuitions.
- → Consequently, Halpern does not have a strong case for the full employment theory with respect to maximum entropy. This supports our larger project of a carefully specified epistemological primacy of information theory in matters of uncertainty.

The Judy Benjamin Case I



The Judy Benjamin Case II

- (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 on Red Second Company ground equals the probability of being on Red Headquarters ground.
- (HDQ) Headquarters inform Judy that in case she is in Red territory, her chance of being on their Headquarters ground is three times the chance of being on their Second Company ground.

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

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

Bayesian Updating

Why Bayesian Updating might not work:

- Bayesian updating requires that an event $T \subseteq \Omega$ is true.
- Sometimes, however, the information received assigns a probability to T or a value to the expectation of a random variable. We will see that the former case can always be reduced to the latter.
- Grove and Halpern try to save Bayesian Updating for the Judy Benjamin case by setting $T =$"receiving information HDQ."

Two Intuitions

(T1) HDQ should not affect $P(A_3)$. Let's call P the prior probability distribution, before receiving HDQ, and Q the posterior probability distribution, after receiving HDQ. Then $Q(A_3) = P(A_3)$.

(T2) If the value of q approaches 1 then $Q(A_3)$ should approach 2/3. HDQ would then be "if you are in Red territory you are almost certainly on Red Headquarters ground." Considering MAP, $Q(A_3)$ should approach 2/3. Continuity considerations pose a contradiction to T1.

Two Intuitions

(T1) HDQ should not affect $P(A_3)$. Let's call P the prior probability distribution, before receiving HDQ, and Q the posterior probability distribution, after receiving HDQ. Then $Q(A_3) = P(A_3)$.

(T2) If the value of q approaches 1 then $Q(A_3)$ should approach $2/3$. HDQ would then be "if you are in Red territory you are almost certainly on Red Headquarters ground." Considering MAP, $Q(A_3)$ should approach $2/3$. Continuity considerations pose a contradiction to T1.

MAXENT and MINXENT I

Depending on whether we have a prior probability distribution we can use MAXENT or MINXENT to calculate the posterior probability distribution using maximum entropy.

MAXENT Choose the probability distribution that fulfills the constraint provided by the information you have and is otherwise maximally uncertain (maximum entropy).

MINXENT Choose the probability distribution that fulfills the constraint provided by the information you have and is otherwise minimally discriminate from the given prior probability distribution (minimum cross-entropy).

MAXENT and MINXENT II

It is most expedient to provide constraints in the form of expectations. We define a payoff function r and then set the expectation to a value α. To be slightly more general, we shall call our events x_i instead of A_i, and $m = 3$. Let $p_i = P(A_i)$ and $q_i = Q(A_i)$.

$$r(x_1) = 2, r(x_2) = 2, r(x_3) = 0, \alpha = \sum_{i=1}^{m} r(x_i)q_i = 1 \qquad \text{(MAP)}$$

$$r(x_1) = 0, r(x_2) = 1, r(x_3) = q, \alpha = \sum_{i=1}^{m} r(x_i)q_i = q \qquad \text{(HDQ)}$$

MAXENT and MINXENT III

For MAXENT, we want to maximize the Shannon entropy:

$$H(Q) = -\sum_{i=1}^{m} q_i \ln q_i \tag{1}$$

For MINXENT, we want to minimize the Kullback-Leibler divergence, given first in its general form as a Lebesgue integral (μ is any background measure for which the Radon-Nikodym derivatives of Q and P exist), then more applicably to the Judy Benjamin case as a finite sum:

$$D(Q||P) = \int_{\Omega} q \ln \frac{q}{p} d\mu$$

$$D(Q||P) = \sum_{i=1}^{m} q_i \ln \frac{q_i}{p_i} d\mu \tag{2}$$

The Constraint Rule for MAXENT I

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

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

There are two constraints for f:

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

$$\text{maximize} \; -\sum_{i=1}^{m} f(x_i) \ln(x_i) \tag{5}$$

The Constraint Rule for MAXENT II

We use Lagrange multipliers to define the functional

$$J(f) = -\sum_{i=1}^{m} f(x_i) \ln f(x_i) + \lambda_0 \sum_{i=1}^{m} f + \lambda_1 \sum_{i=1}^{m} r(x_1) f(x_i) \qquad (6)$$

and differentiate it with respect to x_i

$$\frac{\partial J}{\partial f(x_i)} = -\ln(f(x_i)) - 1 + \lambda_0 + \lambda_1 r(x_i) \qquad (7)$$

The Constraint Rule for MAXENT III

Set (7) to 0 to find the necessary condition to maximize (5)

$$g(x_i) = e^{\lambda_0 - 1 + \lambda_1 r(x_1)} \tag{8}$$

This is the Gibbs distribution. We need to (a) show that the entropy of g is maximal, and (b) show how to find λ_0 and λ_1. (a) is part of the proof, but we won't need it later on. We will only show (b).

The Constraint Rule for MAXENT IV

Let

$$\lambda_1 = -\beta \tag{9}$$

$$Z(\beta) = \sum_{i=1}^{m} e^{-\beta r(x_i)} \tag{10}$$

$$\lambda_0 = 1 - Z(\beta) \tag{11}$$

To find λ_0 and λ_1 we set

$$-\frac{\partial}{\partial \beta} \ln(Z(\beta)) = \alpha \tag{12}$$

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 (3) and (4) are also fulfilled.
First, we show

$$\sum_{i=1}^{m} g(x_i) = \sum_{i=1}^{m} e^{\lambda_0 - 1 + \lambda_1 r(x_i)} = e^{\lambda_0} \sum_{i=1}^{m} e^{\lambda_1 r(x_i)} =$$
$$e^{-\ln(Z(\beta))} Z(\beta) = 1 \tag{13}$$

The Constraint Rule for MAXENT VI

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

$$\alpha = -\frac{\partial}{\partial\beta}\ln(Z(\beta)) = -\frac{1}{\sum_{i=1}^{m} x^{r(x_i)}}\left(\sum_{i=1}^{m} r(x_i)x^{r(x_i)-1}\right)(-x) = \frac{\sum_{i=1}^{m} r(x_i)x^{r(x_i)}}{\sum_{i=1}^{m} x^{r(x_i)}} \tag{14}$$

And, finally,

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

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

$$\alpha Z(\beta)e^{\lambda_0 - 1} = \alpha Z(\beta))e^{-\ln(Z(\beta))} = \alpha \tag{15}$$

The Constraint Rule Applied to Judy Benjamin I

The normalized odds vector for the Judy Benjamin case without using information (HDQ) is:

$$v_0 = (.25, .25, .5)$$

Let's apply the constraint rule to the Judy Benjamin case. Instead of proving an analogous procedure for MINXENT, we first calculate the posterior using the information for (HDQ), but not yet for (MAP), so that we can use the procedure for MAXENT that we just spelled out. (3) is defined by (HDQ). The lambdas are:

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

The Constraint Rule Applied to Judy Benjamin II

We combine the normalized odds vector following from these lambdas ($v_2 = (.16, .48, .36)$) using Dempster's Rule of Combination with (MAP) and get

$$v_1 = (.12, .35, .53)$$

You can check this result by differentiating and examining the critical points of the Kullback-Leibler divergence directly (see next slide). It contradicts intuition (T1) because $q_3 = .53$ is no longer $1/2$.

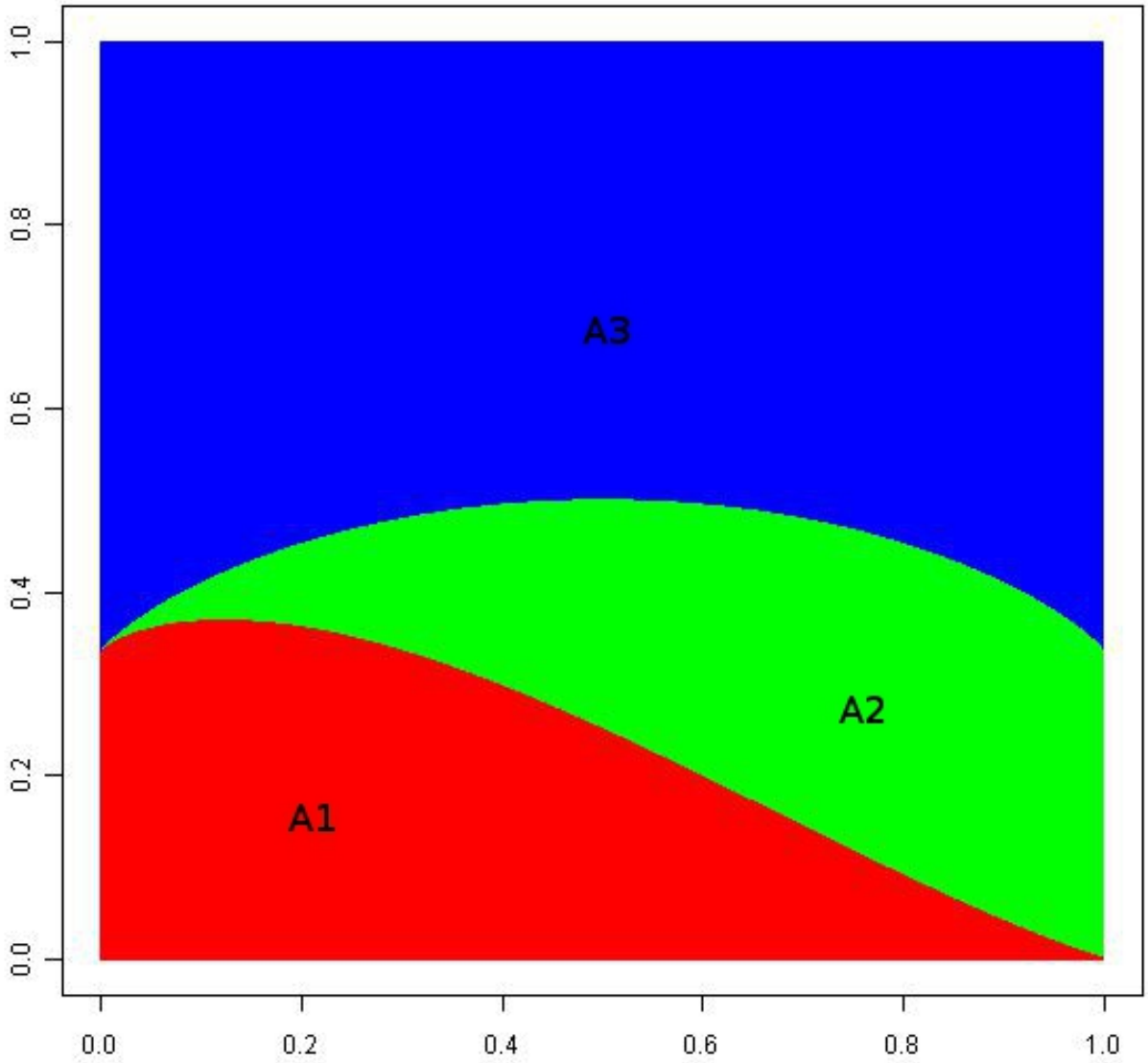
q_1 as a Function of q

Differentiating and examining the critical points of the Kullback-Leibler divergence directly gives us the following function $q \to q_1$:

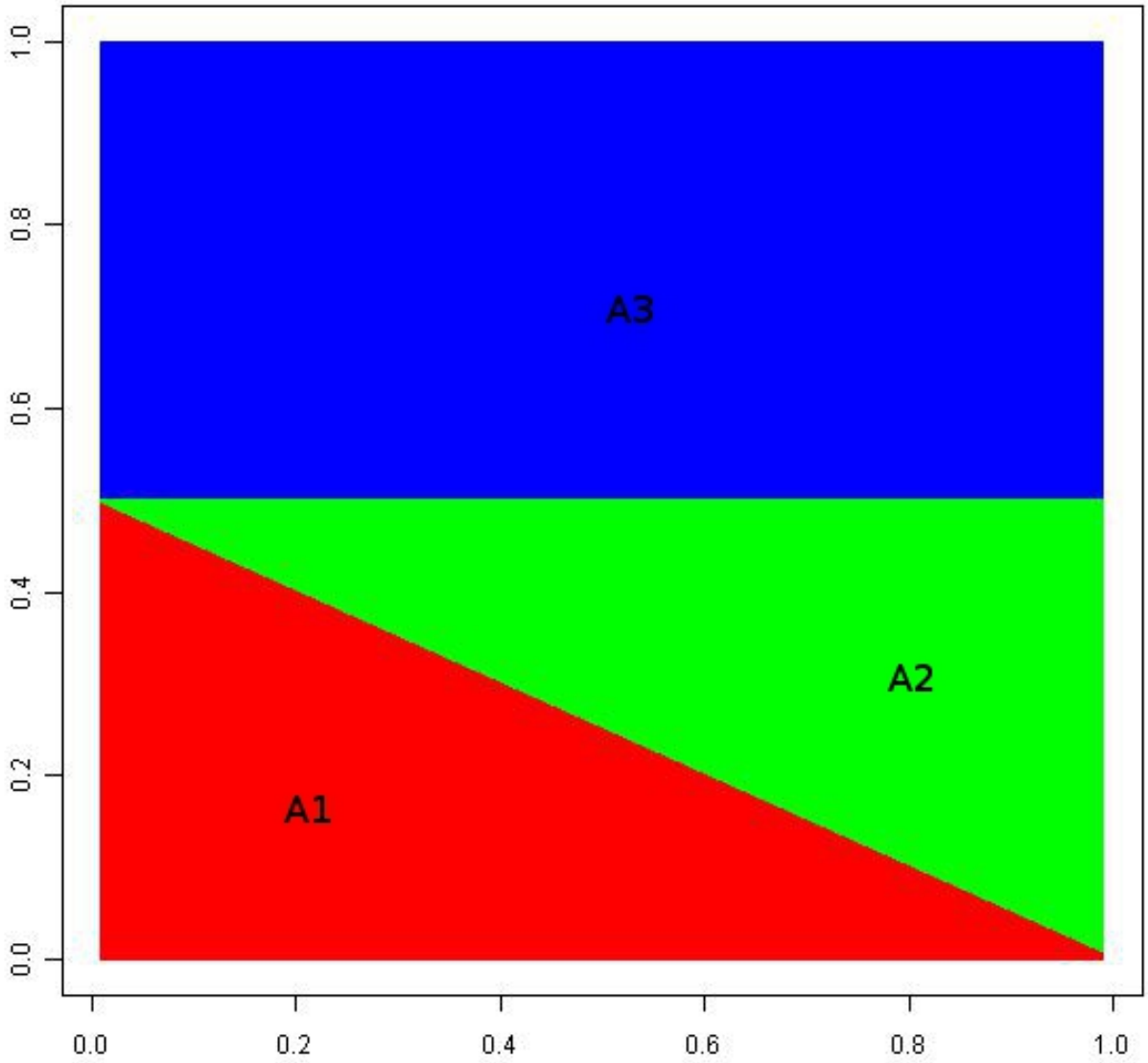
$$q_1 = \frac{s}{1 + st + s}$$

$$s = 2^{-\frac{t \log_2 t + t + 1}{t+1}} \qquad t = \frac{q}{q-1}$$

Intuition (T2)



Intuition (T1)



Four Scenarios I

Here are four scenarios, in which Judy may have received (HDQ):

S1 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 on Headquarters ground, otherwise on Second Company ground. Judy's headquarters find out that the second coin was biased $q : 1 - q$ toward H. The normalized odds vector is $v = (.125, .375, .5)$ and agrees with (T1), because the choice of Blue or Red is completely independent from the choice of Headquarters or Second Company.

Four Scenarios II

S2 Judy's Headquarters has divided the map into 24 congruent rectangles, A_3 into twelve, and A_1 and A_2 into six rectangles each. 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. Thus they relay (HDQ) to Judy, which she correctly interprets with the normalized odds vector $v = (.108, .324, .568)$ (evaluating the 16777216 subsets).

Four Scenarios III-IV

S3 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) randomly selected quadrant to repeat the procedure. Headquarters find out, however, that for A_1, the pilot requires a six to drop off Judy, not just an even number. Thus they relay (HDQ) to Judy, which she correctly interprets with the normalized odds vector $v = (.1, .3, .6)$.

S4 Judy's headquarters know that Judy is not in A_3 and that her chance of being in A_2 is three times the chance of being in A_1. They only succeed, however, in informing Judy of the second part of the message. If Judy had all the information, her normalized odds vector would be $v = (.33, .67, 0)$.

Knowledge Expands Ignorance

In short, given only (MAP) and (HDQ), Judy cannot have any confidence that they are independent. Consequently, she should adjust her probabilities using v_1 in accordance with (T2). Van Fraassen is worried that she has now greater confidence in being in Blue territory, without receiving any specific information about it. As so often, there is discomfort with the ability of maximum entropy to articulate a hypothesis on the basis of ignorance rather than positive evidence. It may be intuitive that an increase in knowledge and certainty about part of the event space also increases the posterior probability of the space about which we are ignorant.

End of Presentation