Probability Kinematics and Halpern's Full Employment Theory

Stefan Lukits

May 27, 2011

1. Introduction

Describe Halpern's full employment theorem and argue that Halpern does little to justify it with respect to MAXENT and MINXENT.

2. Judy Benjamin

There are two conflicting intuitions for the Judy Benjamin case. (I1) When Judy Benjamin learns that $Q(A_1|A_1 \cup A_2) = .75$ (or $q = .75$) it should not affect the probability she assigns to being in blue territory, i.e. $Q(A_3) = P(A_3)$. (For now, our understanding is that P is the prior probability distribution and Q is the posterior probability distribution relative to the information from her headquarters. We also abbreviate $q_1 = Q(A_1), q_2 = Q(A_2), q_3 = Q(A_3)$.) (I2) If her headquarters tell her that $q = 1 - \varepsilon$ with ε a small number, we would expect $q_1 \to 0, q_2 \to 1/3, q_3 \to 2/3$. (I1) and (I2) are contradictory if the map $G : q \to q_1$ is to be continuous.

Giving us very little detail, Bas van Fraassen uses Dempster's Rule of Combination to combine the prior probability distribution with the principle of maximum entropy (MAXENT) to provide the posterior probability distribution suggested by the principle of minimal discrimination (sometimes called MINXENT in analogy to MAXENT). van Fraassen does not mention that he is using Dempster's Rule, nor does he mention that the Rule combined with MAXENT provides us with the result required by MINXENT. His assumptions are correct, however, and Judy Benjamin's posterior probability distribution required by MINXENT is approximately

$$v_1 = (.12, .35, .53) \tag{1}$$

(v_1 is, in van Fraassen's terminology, the normalized odds vector $(Q(A_i))_{i=1,\ldots,m}$. v_0 is the normalized odds vector for P, i.e. $v_0 = (.25, .25, .5)$.) To fill out van Fraassen's description of the Judy Benjamin case and make further observations about (I1) and (I2), we will provide in the next section a proof of the constraint rule and then use the constraint rule to sort out our intuitions.

3. The Constraint Rule

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

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

Because f is a probability distribution it fulfills

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

We want to maximize the 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

$$J(f) = -\sum_{i=1}^{m} f(x_i) \ln f(x_i) + \lambda_0 \sum_{i=1}^{m} f + \lambda_1 \sum_{i=1}^{m} r(x_1) f(x_i) \tag{5}$$

and differentiate it with respect to x_i

$$\frac{\partial J}{\partial f(x_i)} = -\ln(f(x_i)) - 1 + \lambda_0 + \lambda_1 r(x_i) \tag{6}$$

Set (6) to 0 to find the necessary condition to maximize (4)

$$g(x_i) = e^{\lambda_0 - 1 + \lambda_1 r(x_i)} \tag{7}$$

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 [2] and there is no reason to

copy it here. I was not able to find (b) in the literature and will show it here. A faint suggestion on how to do it is in Jaynes' treatment of the Brandeis Dice Problem (see [7, 243]).

Let

$$\lambda_1 = -\beta \tag{8}$$

$$Z(\beta) = \sum_{i=1}^{m} e^{-\beta r(x_i)} \tag{9}$$

$$\lambda_0 = 1 - \ln(Z(\beta)) \tag{10}$$

To find λ_0 and λ_1 we introduce the constraint

$$-\frac{\partial}{\partial\beta} \ln(Z(\beta)) = \alpha \tag{11}$$

To see how this constraint gives us λ_0 and λ_1 Jaynes' solution of the Brandeis Dice Problem 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

$$\sum_{i=1}^{m} g(x_i) = \sum_{i=1}^{m} e^{\lambda_0 - 1 + \lambda_1 r(x_i)} = e^{\lambda_0} \sum_{i=1}^{m} e^{\lambda_1 r(x_i)} = e^{-\ln(Z(\beta))} Z(\beta) = 1 \tag{12}$$

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

$$\alpha = -\frac{\partial}{\partial\beta} \ln(Z(\beta)) = -\frac{1}{\sum_{i=1}^{m} x^{r(x_i)}} \left(\sum_{i=1}^{m} r(x_i) x^{r(x_i)-1} \right) (-x) = \frac{\sum_{i=1}^{m} r(x_i) x^{r(x_i)}}{\sum_{i=1}^{m} x^{r(x_i)}} \tag{13}$$

And, finally,

$$\sum_{i=1}^{m} r(x_i)g(x_i) = \sum_{i=1}^{m} r(x_i)e^{\lambda_0 - 1 + \lambda_1 r(x_1)} = e^{\lambda_0 - 1} \sum_{i=1}^{m} r(x_i)e^{\lambda_1 r(x_1)} =$$
$$e^{\lambda_0 - 1} \sum_{i=1}^{m} r(x_i)x^{r(x_i)} = \alpha e^{\lambda_0 - 1} \sum_{i=1}^{m} x^{r(x_i)} = \alpha e^{\lambda_0 - 1} \sum_{i=1}^{m} e^{-\beta r(x_i)} =$$
$$\alpha Z(\beta)e^{\lambda_0 - 1} = \alpha Z(\beta))e^{-\ln(Z(\beta))} = \alpha \tag{14}$$

4. The Full Employment Theorem

The Kullback-Leibler Divergence is not a metric: it is not symmetric and does not obey the triangle inequality. (Point out Uffink's average mistake.) MINXENT requires that Q be the probability distribution fulfilling the constraint (2) and minimizing the Kullback-Leibler Divergence $D(Q, P)$.

$$D(Q, P) = \sum_{i=1}^{m} q_i \log \frac{q_i}{p_i} \tag{15}$$

We can either use the constraint rule to calculate v_1 (see Green Book 141 and 143) or differentiate $D(Q, P)$ directly for its critical points and make G explicit (see Green Book 148).

$$G(q) = q_1 = \frac{C}{1 + Ct + C} \tag{16}$$

where

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

and

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

In his (somewhat) recent book, *Reasoning About Uncertainty*, Joseph Halpern writes:

> 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. (Halpern [6, 110])

The principle example for this counterintuitive behaviour in Halpern (for another one treated in [11], section 5, see [3]) is the Judy Benjamin case in van Fraassen [12]. Halpern has some personal investment in this case, as he wrote a paper defending a more classical Bayesian approach to the case (see [5]). In this paper, Grove and Halpern want to save intuition (I1) with assumptions that guarantee the independence of $P_{\text{HQ}}(A_3)$ and $P_{\text{HQ}}(A_2|A_1 \cup A_2)$ (the superscript HQ stands for headquarters and emphasizes that Judy Benjamin conditions on the information received by headquarters without necessarily making headquarters' beliefs her own). Indeed, as they set it up, q_3 remains at $1/2$, contrary to (I2), the normalized odds vector v_1 and the function G we just provided.

Both in the paper and in the book, the Judy Benjamin case is a special case in the more general argument that "the only right way to update, especially in an incompletely specified situation, is to think very carefully about the real nature and origin of the information we receive" [5, 3]. Although they do not use the terminology, Grove and Halpern advocate a type of Full Employment Theorem (used in computer science, where due to Turing's halting problem the computer scientist cannot be dispensed with in the writing of computer programs). They end their paper with the call that "one must always think carefully about precisely what the information means" [5, 6].

Halpern's book is a more large-scale presentation of pluralism in reasoning about uncertainty. Possible worlds, probability measures, lower and upper probabilities, Dempster-Shafer belief functions, possibility measures, ranking functions, relative likelihoods, and plausibility measures are introduced, concluding that it is in the end up to the inquirer to choose a representation of uncertainty that fits the bill: "there is no escaping the need to understand the details of the application" [6, 423].

This contrasts with physical statistics where MAXENT is often considered to be a uniquely consistent method of integrating ignorance and information. The uniqueness claim, for example advocated in Shore and Johnson [8], is

contested (see, for example, Uffink [10], but also van Fraassen in [13]), where the authors suggest MAXENT to be a special instance of a family of principles which are consistent relative to specified assumptions). It is widely used in physics, where much effort has gone into showing the consistency of MAXENT and Bayesian updating (see [1], where the family of Renyi functions proposed by Uffink are ruled out, and [4], where the author seeks to "show that MAXENT is capable of producing every aspect of orthodox Bayesian inference and prove the complete compatibility of Bayesian and entropy methods"). (For MAXENT in peaceful coexistence with, conflict with, or as a generalization of Bayesian conditionalization see [11], for Bayesian conditionalization as a generalization of MAXENT see [9].)

We return to Judy Benjamin. Applying MINXENT, we found intuition (I2) confirmed (van Fraassen). As we pointed out at the beginning of this section, however, the Kullback-Leibler Divergence is not symmetric. We are better off calling it the divergence of Q from P, rather than the distance of Q and P in terms of information. Symmetric expressions of relative entropy exist, but they do not have some of the attractive properties of the Kullback-Leibler Divergence (in fact, Kullback and Leibler themselves suggested the symmetric relative entropy $D(P,Q)+D(Q,P)$). In statistics, this asymmetry is often less problematic because the divergence is understood to be *from* a model *to* reality. The asymmetry is justified by the different nature of Q (the model) and P (reality).

Is the asymmetry justified in Judy Benjamin's case? Judy's prior probability distribution is expressed in the normalized odds vector $v_0 = (.25, .25, .5)$. Then she receives the information from headquarters that $Q(A_1|A_1 \cup A_2) = .75$. But perhaps she proceeds in reverse order: first she learns that the chance of being in A_2 is three times the chance of being in A_1 from headquarters, without a prior conception of how large A_1, A_2, A_3 are. She then uses MAXENT rather than MINXENT to establish her first probability distribution conditioned on the information she received from her headquarters:

$$v_2 = (.16, .48, .36) \tag{17}$$

To achieve this result we use the constraint rule with the constraint $r_1 = 0, r_2 = 1, r_3 = q$ (this is equivalent to $q_1 = 3q_2$ in Judy Benjamin's case where $q = 3/4$). (Reader beware, whenever we use the constraint rule there are some hefty calculations.)

$$\lambda_0 = e^{-\frac{q}{1-q}} + e^{-\frac{q^2}{1-q}} \text{ and } \lambda_1 = -\frac{q}{1-q} \tag{18}$$

Notice that the third element of v_2 is already $> 1/3$, in keeping with intuition (I2). Only now Judy Benjamin considers the proportion of the probabilities corresponding to the size of the A_i (A_3 is twice the size of A_1 and A_2). To use the constraint rule, we like to express our constraints in terms of the expectation $\sum_{i=1}^{m} r(x_i)q(x_i)$. In this case, we find that $r_1 = 2, r_2 = 2, r_3 = 0$ is equivalent to the constraint just formulated. Again, using the constraint rule

$$\lambda_0 = 1 - \ln 2 \text{ and } \lambda_1 = -\frac{1}{2}\ln 2 \tag{19}$$

This provides us with the normalized odds vector $v_3 = (.125, .375, .5)$, which is just what (I1) requires! It is a case of which of the two pieces of information we put first whether (I1) or (I2) prevails. If we change $q = 3/4$ to a variable we have $v_3 = (.5(1-q), .5q, .5)$. The continuity required by (I2) is also fully restored. The results for the Judy Benjamin case are not at all counterintuitive as Halpern suggests, but as in the case of Bayesian conditionalization the sequence matters by which we use the information we have.

5. Some Open Questions

Why does Dempster's Rule of Combination give us MINXENT with respect to $D(Q, P)$, not $D(P, Q)$? Dempster's Rule of Combination is symmetrical and presumes that the two pieces of information are independent. Next: give an example of why sequence matters in Bayesian updating, and show that it is not counterintuitive. Address Halpern and Grove's assumptions which favour (I1). Address Friedman and Shimony's counterexample to MAXENT. What is the matter with the graph for G? Label the two pieces of information *map* and *hq*.

[1] Caticha, A. and Giffin, A., 2006. Updating Probabilities. In *MaxEnt 2006, the 26th International Workshop on Bayesian Inference and Maximum Entropy Methods*.

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

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

[4] Giffin, A., 2008. *Maximum Entropy: The Universal Method for Inference*. PhD dissertation, University at Albany, State University of New York, Department of Physics.

[5] 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.

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

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

[8] Shore, J. and 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.

[9] Skyrms, B., 1985. Maximum Entropy Inference as a Special Case of Conditionalization. *Synthese*, 63(1):55–74.

[10] 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.

[11] 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.

[12] 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.

[13] 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.