A Note on the Judy Benjamin Problem

Stefan Lukits

June 24, 2011

Grove and Halpern [1] make a claim that, given the Judy Benjamin problem introduced in Van Fraassen [2], it is not necessary and produces unintuitive consequences to employ the principle of maximum entropy (from now on MAXENT) in order to achieve a reasonable posterior probability distribution. The authors maintain that despite the appearance of partial information (which may be interpreted to call for the use of maximum entropy rather than Bayesian conditionalization), they can handle the problem using Bayes' formula and produce results that are both intuitive and contingent on information not provided in the problem, as they should be because Judy Benjamin's information is so vague.

If we use the constraint rule and MAXENT, Judy Benjamin's posterior probability of being in blue territory is greater than its prior probability. The normalized odds vectors of the prior and posterior probability distribution are, respectively:

$$v_0 = (.25, .25, .5) \tag{1}$$

$$v_1 = (.12, .35, .53) \tag{2}$$

Grove and Halpern consider this development unintuitive, because they think that Judy Benjamin's information given to her by her headquarters is independent of the probability of being in blue territory. I want to show that Grove and Halpern's assumption of independence is unwarranted, because it imposes information on Judy Benjamin's situation that she does not have. I will provide scenarios where independence is given and other scenarios where it is not. The scenarios are not far-fetched, and so we conclude that independence must not be one of our assumptions.

Grove and Halpern claim that it would be natural for Judy Benjamin to

assume independence and uniformity in the probability assignments of her headquarters (they explicitly state, however, that these assumptions go far beyond the information provided in the problem). In contrast to Grove and Halpern, we will look at a scenario where Judy Benjamin makes no assumptions about independence and much weaker assumptions than uniformity by considering measurable partitions of the event space whose grain goes to infinity.

As her partitions become more fine-grained, her posterior probability approaches the posterior probability suggested by MAXENT. Thus, we say, it is true that there is a lot of information Judy Benjamin does not have. She still needs to make a reasonable posterior probability assessment, and we maintain that it is MAXENT which provides it, partly because it is so good at incorporating ignorance. The results will be close to the results suggested by independence, but because independence is not one of our assumptions we have to pay attention to the scenarios in which it is not true and accordingly increase the probability of the event we know less about. The fine-grained partition scenario confirms that this procedure is in full accordance with intuitions.

Here are three scenarios, in which Judy may have received her information:

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 with $q = .75$. 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.

S2 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 (HQ) to Judy, which she correctly interprets with the normalized odds vector $v = (.1, .3, .6)$.

S3 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 (HQ) to Judy, which she correctly interprets with the normalized odds vector $v = (.108, .324, .568)$ (evaluating the 16777216 subsets).

Scenarios S1 and S2 are good examples of where independence is true and where it is not, respectively. In the following section, we will focus on S3 and consider what happens when the grain of the partition becomes finer. To do this, let

$$A_1 \cup A_2 \cup A_3 = A = \bigcup_{i \in I} B_i \text{ and } \mathcal{A} = 2^{\{B_i\}_{i \in I}} \tag{3}$$

where I is a finite set of indices and the B_i are a pairwise disjoint covering of A with $B_i \cap B_j = \emptyset$ for $i \neq j$ as well as $\mu(B_i) = \mu(B_j)$ for all $i, j \in I$ and an appropriate measure μ.

Now let $\mathcal{B} \subset \mathcal{A}$ be the set that contains only those collections of B_i for which there are $t = q/(1-q)$ (in Judy Benjamin's case, $t = 3$) times as many B_i in A_2 as there are in A_1. In other words, $B \in \mathcal{B}$ iff

$$\text{and } q\mu\left(\bigcup_{i \in J \subset I} B_i \cap A_1\right) = (1-q)\mu\left(\bigcup_{i \in J \subset I} B_i \cap A_2\right) \tag{4}$$

For simplicity's sake, we assume that $\#J = 4n = 4ts$ (where t depends on q as above and s indicates the grain of the partition). (Show a graph with the example $t = 3$ and $s = 2$.) Brute combinatorics tells us that $H_3(s,t)$, the average ratio of $\mu(B \cap A_3)/\mu(B)$ for all $B \in \mathcal{B}$ is

$$H_3(s,t) = \frac{1}{N} \sum_{j=0}^{2ts} \sum_{i=0}^{s} \binom{2ts}{j} \binom{ts}{i} \binom{ts}{ti} \varphi_{ij} \tag{5}$$

where

$$N = 2^{2ts} \sum_{i=0}^{s} \binom{ts}{i} \binom{ts}{ti} \tag{6}$$

and

$$\varphi_{ij} = j(j + i(1+t))^{-1} \tag{7}$$

This would be the proper posterior $Q(A_3)$ for scenario S3 with $t = 3$ and $s = 2$. We are interested in what happens to $H_3(s,t)$ as $s \to \infty$ (the partition becomes more fine-grained; (T1) suggests that $H_3(s,t)$ will remain at approximately $1/2$, while (T2) suggests that $H_3(s,t)$ is consistently larger than $1/2$ and more so with greater t) and as $t \to \infty$ (in this case, (T2) suggests that $H_3(s,t)$ should approach $2/3$, while (T1) suggest that it should approach $1/2$).

Here is a table with the results for $t = 1, \ldots, 10$ and $s = 1, \ldots, 25$:

	1	2	3	4	5	6	7	8	9	10
1	0.2917	0.3732	0.4094	0.4297	0.4426	0.4515	0.4580	0.4629	0.4669	0.4700
2	0.4327	0.5187	0.5682	0.5976	0.6143	0.6241	0.6305	0.6351	0.6386	0.6413
3	0.4812	0.5271	0.5512	0.5635	0.5705	0.5751	0.5784	0.5809	0.5829	0.5845
4	0.4958	0.5274	0.5539	0.5734	0.5893	0.6028	0.6141	0.6233	0.6306	0.6363
5	0.4997	0.5269	0.5531	0.5711	0.5834	0.5917	0.5973	0.6012	0.6040	0.6062
6	0.5006	0.5265	0.5526	0.5705	0.5832	0.5930	0.6013	0.6085	0.6151	0.6212
7	0.5006	0.5263	0.5523	0.5705	0.5837	0.5937	0.6013	0.6070	0.6114	0.6147
8	0.5005	0.5261	0.5521	0.5702	0.5831	0.5925	0.5998	0.6056	0.6106	0.6150
9	0.5004	0.5259	0.5520	0.5701	0.5831	0.5930	0.6008	0.6071	0.6123	0.6166
10	0.5004	0.5258	0.5519	0.5700	0.5830	0.5927	0.6001	0.6059	0.6105	0.6144
11	0.5003	0.5257	0.5518	0.5699	0.5829	0.5926	0.6002	0.6064	0.6116	0.6161
12	0.5002	0.5256	0.5517	0.5698	0.5828	0.5926	0.6001	0.6061	0.6110	0.6149
13	0.5002	0.5256	0.5516	0.5698	0.5828	0.5925	0.6001	0.6061	0.6112	0.6155
14	0.5002	0.5255	0.5516	0.5697	0.5827	0.5925	0.6000	0.6061	0.6111	0.6152
15	0.5002	0.5255	0.5515	0.5697	0.5827	0.5924	0.6000	0.6060	0.6110	0.6152
16	0.5001	0.5254	0.5515	0.5696	0.5826	0.5924	0.6000	0.6060	0.6110	0.6152
17	0.5001	0.5254	0.5514	0.5696	0.5826	0.5923	0.5999	0.6060	0.6110	0.6151
18	0.5001	0.5254	0.5514	0.5695	0.5826	0.5923	0.5999	0.6060	0.6110	0.6152
19	0.5001	0.5254	0.5514	0.5695	0.5825	0.5923	0.5999	0.6059	0.6109	0.6151
20	0.5001	0.5253	0.5513	0.5695	0.5825	0.5923	0.5998	0.6059	0.6109	0.6151
21	0.5001	0.5253	0.5513	0.5695	0.5825	0.5922	0.5998	0.6059	0.6109	0.6151
22	0.5001	0.5253	0.5513	0.5694	0.5825	0.5922	0.5998	0.6059	0.6109	0.6151
23	0.5001	0.5253	0.5513	0.5694	0.5824	0.5922	0.5998	0.6059	0.6109	0.6150
24	0.5001	0.5253	0.5512	0.5694	0.5824	0.5922	0.5998	0.6059	0.6109	0.6150
25	0.5001	0.5252	0.5512	0.5694	0.5824	0.5922	0.5998	0.6059	0.6108	0.6150

It appears that (T2) is in this case the better intuition to follow. The appearance is confirmed by the following consideration. Let X be the random

variable

$$X = \frac{\mu(B \cap A_3)}{\mu(B)} \tag{8}$$

for a randomly chosen $B \in \mathcal{B}$. Let $X_i = \mu(B \cap A_i)$ also be random variables for $i = 1, 2, 3$. Then the expectation of X is

$$EX = \frac{EX_3}{\sum_{i=1}^{3} EX_i} \tag{9}$$

We know that according to the Moivre-Laplace Theorem and continuity correction (basically approximating the binomial distribution, which is difficult to calculate for large integers, by the normal distribution)

$$\sum_{k=0}^{m} \binom{n}{k} \approx 2^n \int_{-\infty}^{m+\frac{1}{2}} N\left(\frac{n}{2}, \frac{n}{4}\right)(x)dx \tag{10}$$

and

$$\sum_{k=0}^{m} k\binom{n}{k} \approx 2^n \int_{-\infty}^{m+\frac{1}{2}} xN\left(\frac{n}{2}, \frac{n}{4}\right)(x)dx \tag{11}$$

with

$$N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{12}$$

As above, combinatorics show that

$$EX_1 = \frac{\sum_{i=0}^{s} i\binom{ts}{i}}{\sum_{i=0}^{s} \binom{ts}{i}} \tag{13}$$

$$EX_2 = \frac{\sum_{i=0}^{ts} i\binom{ts}{i}}{\sum_{i=0}^{ts} \binom{ts}{i}} \tag{14}$$

$$EX_3 = \frac{\sum_{i=0}^{2ts} i\binom{2ts}{i}}{\sum_{i=0}^{2ts} \binom{2ts}{i}} \tag{15}$$

Consequently,

$$EX_1 \approx \frac{\int_{-\infty}^{s+\frac{1}{2}} xN\left(\frac{n}{2}, \frac{n}{4}\right)(x)dx}{\int_{-\infty}^{s+\frac{1}{2}} N\left(\frac{n}{2}, \frac{n}{4}\right)(x)dx} \tag{16}$$

Using the well-known integrals (erf is the Gauss error function)

$$\int_a^b \omega e^{-\omega^2} d\omega = \frac{1}{2}e^{-a^2} - \frac{1}{2}e^{-b^2} \tag{17}$$

and

$$\int_a^b e^{-\omega^2} d\omega = \frac{\sqrt{\pi}}{2} \left(\operatorname{erf}(b) - \operatorname{erf}(a)\right) \tag{18}$$

as well as the substitution

$$y = \sqrt{\frac{2}{n}} \left(x - \frac{n}{2}\right) \tag{19}$$

(16) simplifies to

$$EX_1 = \frac{n}{2} \left(1 - \sqrt{\frac{2}{\pi n}} \cdot \frac{e^{-w_1^2}}{1 + \operatorname{erf}(w_1)}\right) \tag{20}$$

with (remember that $n = ts$)

$$w_1 = \sqrt{\frac{2}{n}} \left(s + \frac{1}{2}\left(1 - n\right)\right) \tag{21}$$

Analogously, EX_2 and EX_3 approximately simplify to (respectively)

$$EX_2 = \frac{n}{2} \left(1 - \sqrt{\frac{2}{\pi n}} \cdot \frac{e^{-w_2^2}}{1 + \operatorname{erf}(w_2)}\right) \tag{22}$$

$$EX_3 = n \left(1 - \sqrt{\frac{1}{\pi n}} \cdot \frac{e^{-w_3^2}}{1 + \operatorname{erf}(w_3)}\right) \tag{23}$$

with

$$w_2 = \frac{n+1}{n\sqrt{2}} \qquad w_3 = \frac{2n+1}{2\sqrt{n}} \tag{24}$$

Evaluating these expressions using L'Hôpital's rule it becomes apparent that they show the following behaviour as $s \to \infty$ and t remains fixed:

$$EX_1 \to \frac{s}{2} \left(t - \sqrt{2} \cdot \frac{t^2 - 4}{t}\right) \tag{25}$$

$$EX_2 \to \frac{n}{2} \tag{26}$$

$$EX_3 \to n \tag{27}$$

X_2 and X_3 behave precisely as we would expect them to behave: sloppily speaking, taking a random subset from A_2 or A_3, we would expect it to be

half the size of A_2 or A_3 respectively. (What if t is 1: wouldn't you expect $EX_1 = n/2$ as well?) The constraint that the randomly chosen subset B fulfills (again, expressed sloppily) $B \cap A_2 = t \cdot B \cap A_1$, however, introduces an intriguing new factor for EX_1 that is dependent on the two factors of n, the grain s and the constraint t.

We now have several different ways of calculating q_t, the posterior probability of the event that Judy Benjamin is in A_1. If we want to use a particular partition in the spirit of S3, we need to consult the table provided above. Using either MINXENT or the constraint rule and differentiating the Kullback-Leibler Divergence to achieve the result according to MAXENT, q_t will be value solely dependent on t:

$$G(q) = q_1 = \frac{C}{1 + Ct + C} \tag{28}$$

where

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

and

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

Grove and Halpern suggest that simply

$$q_t = \frac{1}{2}\left(1 - \frac{1}{1 + t}\right) \tag{29}$$

And now our fine grain partition method suggests that q_t should approximately follow equation (9) using the explicit expressions given for EX_i above.

[1] 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.

[2] 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.