REGRESSION TO THE MEAN AND JUDY BENJAMIN

RANDALL G. MCCUTCHEON

Abstract. Van Fraassen's Judy Benjamin problem asks how one ought to update one's credence in A upon receiving evidence of the sort "A may or may not obtain, but B is k times likelier than C", where $\{A, B, C\}$ is a partition. Van Fraassen's solution, in the limiting case $k \to \infty$, recommends a posterior converging to $P(A|A \cup B)$ (where P is one's prior probability function). Grove and Halpern, and more recently Douven and Romeijn, have argued that one ought to leave credence in A unchanged, i.e. fixed at $P(A)$. We argue that while the former approach is superior, it brings about a reflection violation due in part to neglect of a "regression to the mean" phenomenon, whereby when C is eliminated by random evidence that leaves A and B alive, the ratio $P(A) : P(B)$ ought to drift in the direction of $1 : 1$.

1. Introduction

In this paper, we describe a selection effect at work in the Judy Benjamin problem. In a limit case, we characterize this effect as "regression to the mean"–probabilities of surviving cells of a partition trend closer in ratio in response to exposure to unknown evidence that fails to eliminate either. Citing heuristic arguments based on this reasoning, as well as a simulation having suitably generic protocols, we accordingly claim that Judy's posterior credences must, in the limit, lie between those favored by two influential extant positions. Our arguments do not pin down an exact solution, indicating only a rough estimate of what we take to be best policy. Though it follows that we've painted this best policy in somewhat broad strokes, we shall indicate reasons why no precise canonical solution is likely to emerge.

2. Working assumptions

Bas van Fraassen (1981, p. 377) introduces an updating puzzle starring Judy Benjamin (a fictional character from the 1980 film *Private Benjamin*). Judy is involved in a war games exercise. Van Fraassen writes:

> The war games area is divided into the region of the Blue Army, to which Judy Benjamin and her fellow soldiers belong, and that of the Red Army. Each of these regions is further divided into Headquarters Company Area and Second Company Area. The patrol has a map which none of them understands, and they are soon hopelessly lost. Using their radio they are at one point able to contact their own headquarters. After describing whatever they remember of their movements, they are told by the duty officer "I don't know whether or not you have strayed into Red Army territory. But if you have, the probability is $\frac{3}{4}$ that you are in their Headquarters Company Area." At this point the radio gives out.

Van Fraassen now asks:

> *Question:* What will be Private Benjamin's posterior probability that she is in the friendly Blue Army Region?

Using an information distance method, van Fraassen answers $\approx .5327$. Grove and Halpern (henceforth G&H, 1997) and Douven and Romeijn (henceforth D&R, 2011) argue for the "intuitive" answer, i.e. $\frac{1}{2}$.

Before proceeding, we need to get straight on what we take the problem to be. We'll do this by formulating several "working assumptions".

Assumption 1: Judy's prior is[1]

$$\Big(J(Blue\ HQ), J(Blue\ 2nd), J(Red\ HQ), J(Red\ 2nd)\Big) = \Big(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\Big).$$

Assumption 2: Judy regards the duty officer (henceforth HQ) as an expert relative to her.

There is a choice for our next assumption, which regards the protocol for HQ's message to Judy. Denote HQ's credence function by P. G&H (1997) employ:

Assumption $3_{\mathbf{A}}$: HQ reports (always and only) the ratio $P(Red\ HQ) : P(Red\ 2nd)$.

Van Fraassen's original prose formulation suggests a different protocol:[2] "I don't know whether or not you have strayed into Red Army territory" explicitly implies that HQ has not ruled out $Blue$.[3] Hence we'll also consider the alternative:

---

[1]Other possibilities: (1) it is our expectation of Judy's prior that is $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$, and (2) Judy's prior is simply $(J(Blue), J(Red\ HQ), J(Red\ 2nd)) = (\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$ (i.e. suppress subdivision of the $Blue$ region). Although these alternatives might affect proper analyses in interesting ways, neither would materially impact occurrence of the qualitative features that constitute our concern here.

[2]Note however that van Fraassen also gives a "coarsest description" of the problem more along the lines of Assumption $3_{\mathbf{A}}$. We don't know which interpretation he intends his solution to address.

[3]We take $P(Blue) \neq 0$ and "HQ has not ruled out $Blue$" to be equivalent. Also we take HQ to report a ratio $0 : 0$ if and only if he has ruled out Red. Other practices are possible.

Assumption 3_B: HQ reports the ratio $P(Red\ HQ) : P(Red\ 2nd)$ and indicates whether or not $Blue$ has been ruled out (i.e. whether or not $P(Blue) = 0$).[4]

These competing third assumptions can yield different answers: if, after learning that $P(Red\ HQ) : P(Red\ 2nd)$ is $3 : 1$, Judy were to ask HQ "Have you ruled out $Blue$?", she would upon receiving the answer *yes* lower her credence in $Blue$ (to zero). This implies that she would, provided her prior in *yes* were positive, raise her credence in $Blue$ upon receiving the answer *no*. But receiving the answer *no* leaves her in the same situation as in van Fraassen's original prose formulation.

Our interpretation of the above is that the Judy Benjamin problem is a problem in (expert) reflection. (On "reflection", see van Fraassen 1984 and especially Schervish et. al. 2004.) Since HQ is an expert relative to Judy, her credences ought to be the expectation of his. As Judy's story is underspecified, however, this expectation runs "over all possible story completions" weighted in proportion to likelihood, given what we already know. Exact calculation of such an expectation being impractical, the solution therefore turns on making an additional, plausibly *generic* (as used by us, this term means roughly "representative of the mean") stipulation, rule or assumption allowing for direct computation of an expectation.

3. Van Fraassen

Van Fraassen's attempt at a generic updating rule is introduced in this passage:

> Let the agent's prior belief state be characterised by the probability function P and his posterior state by P'. Given any measurable partition X on which P is positive...the relative information in P' with respect to P...is defined by $I(P', P, X) = \sum\{P'(A) \log \frac{P'(A)}{P(A)} : A \in X\}$. The deliverances of experience place a constraint on what the posterior P' should be.... The *Infomin Principle*, as I shall call it, now says that the agent should choose his posterior P' so as to satisfy that constraint while minimizing information relative to P.

[4]Some authors would likely disavow both protocols. D&R (2011), for example, treat HQ's message as a conditional rather than as a report of a conditional probability. We find this implausible. Indeed, we don't think it is obvious which conditional "if you are in the Red area, the probability is $\frac{3}{4}$ that you are in the Red HQ area" is supposed to represent–in particular because we don't know what proposition is intended to serve as consequent. (It can't be "the probability is $\frac{3}{4}$ that you are in the Red HQ area".) One might rephrase to something like "if I were to learn that you are in the Red area then my posterior probability in Red HQ would be $\frac{3}{4}$", but we don't see how this would require analysis different from a conditional probability report. More natural, by far, is to simply assume that HQ's message is a (fairly standard) species of vernacular for such a report.

For Judy, the constraint points to a posterior Q of $(x, \frac{3}{4}(1-x), \frac{1}{4}(1-x))$ for some $x \in [0,1]$, and the Infomin Principle bids her choose x so as to minimize

$$I(Q,P) = x\Big(\log x - \log\frac{1}{2}\Big) + \frac{3}{4}(1-x)\Big(\log(\frac{3}{4}(1-x)) - \log\frac{1}{4}\Big) + \frac{1}{4}(1-x)\Big(\log(\frac{1}{4}(1-x)) - \log\frac{1}{4}\Big).$$

The minimum occurs at $Q(Blue) = x \approx .5327$.

We reject the Infomin treatment on the grounds that it violates reflection. Under protocol $\mathbf{3_A}$ this is obvious. For if Judy learns that the ratio $P(Red\ 2nd) : P(Red\ HQ)$ is $1 : q$, $q > 0$, then Judy's posterior under Infomin is

$$Q(Blue) = \frac{2q^{q/(q+1)}}{2q^{q/(q+1)} + q + 1} = h(q) \geq \frac{1}{2},$$

with equality only at $q = 1$. (The dubiousness of this was noted in Seidenfeld 1986.) Ratio reports not of the form $1 : q$ for $q > 0$ lead to credences $\geq \frac{1}{2}$ as well; van Fraassen maintains that Judy (by continuity) ought to update her credence in $Blue$ to $\frac{2}{3}$ for each of the reported odds ratios $0 : 1$ and $1 : 0$, and obviously to 1 upon report of the indeterminate ratio $0 : 0$. So unless Judy is also learning that $Blue$ has not been eliminated (in which case there is an additional scenario where credence in $Blue$ drops), this is a unidirectional shift in credence, and so a reflection violation.

Infomin violates reflection under $\mathbf{3_B}$ as well. For maximum generality, we will give a proof of this that does not depend on any subdivision of the $Blue$ region. First, we introduce a conservative principle for generic updating in line with the themes we are promoting. For a region R, let e_R be Judy's probability that R is eliminated by HQ, conditional on $\neg R$. The principle asserts that in the generic (no story) case, if L and S are atomic regions with $J(L) \geq J(S)$, then $e_L \leq e_S$. That is, *larger regions are not more likely to be eliminated, should they be non-actual.* More generally, if L (respectively S) is a disjoint union of atomic regions $L_1, \ldots, L_m$ (respectively $S_1, \ldots, S_n$, $n \leq m$) with $J(L_i) \geq J(S_i)$, $1 \leq i \leq n$, one has the same conclusion.

Taking now $Blue$ as an atomic region, let p_0 (respectively p_1, p_2, p_3, p_4, p_5, p_6) be the probability that the set of regions $\mathcal{E}$ eliminated by HQ is $\emptyset$ (respectively $\{Red\ HQ\}$, $\{Red\ 2nd\}$, $\{Blue\}$, $\{Red\ HQ, Red\ 2nd\}$, $\{Red\ HQ, Blue\}$, $\{Red\ 2nd, Blue\}$). By symmetry we take it that $p_1 = p_2$ and $p_5 = p_6$. Note that if $p_0 = 1$ then there's an obvious reflection violation; so we assume $p_0 < 1$.

Applying the former principle to the "smaller, larger" pairs "$Red\ HQ$, $Blue$" and "Red, $\neg Red\ 2nd$" yields the inequalities $p_1 + p_4 + p_5 \geq \frac{3}{2}p_3 + 3p_5$ and $p_4 \geq 2p_5$. Using now the previous identities and the two inequalities in turn, a $\frac{2}{3}$-limiting caser's expectation of $P(Blue)$ conditional on there being at least one region eliminated is

$$\begin{aligned} E\Big(P(Blue)|\mathcal{E} \neq \emptyset\Big) &= \frac{\frac{2}{3}p_1 + \frac{2}{3}p_2 + p_4}{p_1 + p_2 + p_3 + p_4 + p_5 + p_6} = \frac{\frac{4}{3}p_1 + p_4}{2p_1 + p_3 + p_4 + 2p_5} \\ &\geq \frac{p_1 + \frac{1}{2}p_3 + \frac{2}{3}p_4 + \frac{2}{3}p_5}{2p_1 + p_3 + p_4 + 2p_5} \geq \frac{p_1 + \frac{1}{2}p_3 + \frac{1}{2}p_4 + p_5}{2p_1 + p_3 + p_4 + 2p_5} = \frac{1}{2}. \end{aligned}$$

Since Infomin is such and also says $E(P(Blue)|\mathcal{E} = \emptyset) > \frac{1}{2}$, it violates reflection.[5]

4. The one-half solution, part one: Douven and Romeijn

D&R (2011) formulate HQ's message as follows:

> (1) I can't be sure where you are. If you are in Red territory, the odds are 3 : 1 that you are in Headquarters Company area.

It is arguable (at least from considerations of conversational implicature) that HQ communicates, through (1), that he has not eliminated *Blue*. The formulation of the limit case is similar:

> (2) If you are in Red territory, then the odds are 0 : 1 that you are in Second Company area

One might claim that the case is slightly different here–(2) does not contain the phrase "I can't be sure where you are"–but we still think it likely that (2) communicates (conversational implicature again) that *Blue* has not been eliminated. In any event this is the more charitable reading; the general one-half solution D&R argue for has it that Judy never changes her credence in *Blue* in response to a report of the form "If you are in Red territory, the odds are $x : y$ that you are in Headquarters Company area", where at least one of x, y is non-zero. On the other hand, it is clearly the case that Judy will raise her credence in *Blue* to 1 when the ratio 0 : 0 is reported. Assuming that *Red* is eliminated with positive probability, there must then be an additional scenario where Judy's credence in *Blue* drops in order to avoid a reflection violation. Assuming that HQ will make explicit that he has eliminated *Blue* whenever this is true provides the requisite scenario. The upshot is that a critique of D&R may without loss of generality assume that HQ follows protocol $\mathbf{3_B}$, and in particular that (2) imparts implicit evidence that HQ has *not* eliminated *Blue*.

D&R's argument is surprisingly direct–they begin by, in effect, brutely stipulating that $Q(Blue) = \frac{1}{2}$ (by listing it among their "desiderata" for Judy's posterior Q). Against the charge (attributed to Jon Williamson in particular) of begging the question, D&R write "...it is entirely unclear how one could beg any questions simply by registering one's intuitive verdict (as opposed to giving an argument)...". We don't view the move as question-begging, but as a genericity claim fairly based on first blush naturalness considerations.

[5]This argument needn't vitiate the $\frac{2}{3}$ limiting case conclusion in general, although it does show that supporters of such a conclusion must endorse a posterior in *Blue* strictly less than $\frac{1}{2}$ at some middle range(s) of the reported conditional probability $P(Red\ HQ|Red)$. The $\frac{2}{3}$ limiting case conclusion would be apt, in fact, for a variant of the problem in which *Blue* is subdivided into two regions and HQ reports that *both* are live. In this case the effect we are studying manifests at the middle ranges as a "progression from the mean" (Judy raises credence in both *Red* subregions above the mean value $\frac{1}{4}$) rather than a "regression to the mean". We reserve the latter notion ("regression") for cases in which one learns nothing beyond which cells in some partition survive.

D&R buttress their intuitions with a contention that "$\frac{2}{3}$ limiting casers" treat (2) as a material conditional, when they ought instead to treat it as an indicative conditional. They then note the existence of indicative conditionals $A \rightarrow B$, to accept which doesn't cause a change in one's credence in the antecedent A. They give the example 'If it rains tomorrow, we cannot have sundowners at the Westcliff', asserted upon learning that the Westcliff's indoor area will be occupied by a wedding party. For a contrary case, they consider 'If Henry robbed the jeweler, then he also shot him', uttered in a context where one is confident that Henry is no murderer. Here, acceptance of the conditional does causes one's credence in the antecedent to fall.

Our "average over all stories" motif suggests that one should, in the generic case, assume a correlation lying between these extremes. D&R, contrarily, hearken to an intuition that (2) "does not seem to contain any information relevant to whether she is in Red rather than in Blue territory...". The result is that they treat (2) as a conditional lying at one of the two exhibited extremes (that of "sundowners").

The policy implied by this practice leads, however, to inconsistency. Recall, HQ's protocol is to report the ratio $P(Red\ HQ) : P(Red\ 2nd)$ and indicate whether or not $P(Blue) = 0$. Imagine now a counterfactual protocol, according to which HQ reports the ratio $P(Red\ 2nd) : P(Blue)$ and indicates whether or not $P(Red\ HQ) = 0$. Suppose that HQ sends the following message under the counterfactual protocol:

> (2a) If you aren't in the Red Headquarters area, then the odds are 0 : 1 that you are in the Red Second Company area

Their treatment of (2) commits D&R to the intuition that (2a) "does not seem to contain any information relevant to whether she is in Red Headquarters area". That indicates a posterior probability of $\frac{3}{4}$ in $Blue$ upon receipt of (2a). But that indicates, in turn, a posterior probability of $Q(Blue) = \frac{3}{4}$ upon receipt of (2) as well, since HQ's message under the counterfactual protocol will be (2a) in precisely those circumstances that HQ's message under the original protocol will be (2)–namely, when $Red\ 2nd$ is eliminated but $Blue$ and $Red\ HQ$ are not.[6]

The Henry conditional has a story associated with it ensuring that it will induce the behavior it does (we are told to be confident that Henry is no murderer), as does the sundowners conditional (we are told that the Westcliff's indoor area will be occupied by a wedding party). The unembellished (2) is more analogous to something like:

> (3) If Romeo seduces Helena, they'll be married

[6]On pain of inconsistency, anyone who subscribes to the general one-half solution must either deny this seeming truism or claim that $Red\ 2nd$ is not eliminated almost surely. (Drawing *semantic* distinctions between the conditionals (2) and (2a) misses the point, for it is never the semantic content of a received message that is conditioned on, but the fact that it was received.) Grove and Halpern (1997) is the best attempt; we critique their solution in the next section. We concede that (2) and (2a) are assertible with positive probability and equivalent, both to each other and to (2b) *You are not in the Red Second Company area*, asserted by a duty officer whose protocol is to report the value of $P(Red\ 2nd)$ and to indicate so if he knows which region Judy is in. Where we differ from van Fraassen and other $\frac{2}{3}$ limiting casers is in our admission of regression, whereby $Blue$ and $Red\ HQ$ trend closer in probability when $Red\ 2nd$ is eliminated and they are not.

What story goes along with (3) is left to the imagination. On one, (3) is to be interpreted as report of Romeo's honor; he would not seduce a woman he would not also marry. Here acceptance of (3) causes credence in *no seduction* to rise (say from $\frac{1}{2}$ to $\frac{3}{4}$). On another, (3) is to be interpreted as report that Helena's father is standing by, shotgun in hand, bent on seeing any man who seduces his daughter wed to her posthaste. Here acceptance of (3) leaves credence in *no seduction* where it was (say $\frac{1}{2}$). Lacking a story, it appears sensible, contra D&R, to "average over all stories", and adopt policies between these extremes.

5. The one-half solution, part two: Grove and Halpern

Our case in the previous section was based in part on a "seeming truism"–that (2) and (2a) have equivalent assertion conditions under the appropriate protocols–leading to an apparent inconsistency in the one-half solution brought about by two different ways of reporting the elimination of the *Red 2nd* region. Against this, G&H (1997) present a version of the one-half solution that is "almost surely consistent". This is accomplished (in part) by having Judy adopt a model on which the probability of HQ eliminating *Red 2nd* (or either of the other two regions) is zero.

The space of probability functions on the algebra generated by

$$\{Blue\ 2nd, Blue\ HQ, Red\ 2nd, Red\ HQ\}$$

can be identified with the set of quadruples

$$\{(a,b,c,d) : a,b,c,d \geq 0,\ a+b+c+d=1\}.$$

This set forms a regular pyramid (convex hull of the extreme measures $a=1$, $b=1$, $c=1$ and $d=1$) in 4-space. A natural (say G&H) candidate for a generic prior distribution on the location of HQ's probability function in this space is the (standard) uniform one. Their "additional assumption" is thus:

GH: Judy treats HQ's credence function on

$$\{Blue\ HQ, Blue\ 2nd, Red\ HQ, Red\ 2nd\}$$

as a uniformly (with respect to the usual measure) distributed random variable on the convex hull of $(1,0,0,0),(0,1,0,0),(0,0,1,0),(0,0,0,1)\}$.

GH is in agreement with Judy's prior, which, recall, we take to assign probability $\frac{1}{4}$ to each of the four regions. Applying now the change of variables $x=a+b$, $y=c$, $z=d$ and identifying triples $(x,y,z=1-x-y)$ with elements of $R=\{(x,y):0\leq x\leq 1,\ 0\leq y\leq 1-x\}$ gives Judy the prior distribution on HQ's credence function

$$(P(Blue), P(Red\ 2nd), P(Red\ HQ)) = (x,y,1-x-y)$$

determined by the density function $g(x,y)=6x$. Now when HQ transmits the message "$P(Red\ HQ|Red)=\frac{3}{4}$" (an event of measure zero), G&H "coarsen" this message and have Judy condition on "$\frac{3}{4}-\epsilon \leq P(Red\ HQ|Red) \leq \frac{3}{4}+\epsilon$" for some

small $\epsilon > 0$.[7] So Judy's posterior credence in $Blue$ is the expectation of x on

$$\begin{aligned} R_1 =& \Big\{ (x,y) \in R : \frac{3}{4} - \epsilon \leq \frac{y}{y+z} \leq \frac{3}{4} + \epsilon \Big\} \\ =& \Big\{ (x,y) : (\frac{3}{4} - \epsilon)(1-x) \leq y \leq (\frac{3}{4} + \epsilon)(1-x) \Big\} \end{aligned}$$

Namely,

$$Q(Blue) = \frac{\int_{R_1} x(6x)\, dx}{\int_{R_1} 6x\, dx} = \frac{\epsilon}{2\epsilon} = \frac{1}{2}.$$

G&H implicitly maintain that the $\frac{2}{3}$ limiting case argument involves a conflation of the message received, namely $\frac{P(Red\ HQ)}{P(Red)} = 1$, with the alternate message $P(Red\ 2nd) = 0$. (Compare (2) and (2b) of the previous section.) A natural coarsening of the latter is $P(Red\ 2nd) \leq \epsilon$. Letting then

$$R_2 = \{(x,y) \in R : y \geq 1 - x - \epsilon\},$$

this results in what G&H consider the apocryphal limiting case posterior

$$Q_l(Blue) = \lim_{\epsilon \to 0} \left(\frac{\int_{R_2} x(6x)\, dx}{\int_{R_2} 6x\, dx} \right) = \lim_{\epsilon \to 0} \left(\frac{2\epsilon - 3\epsilon^2 + 2\epsilon^3 - \frac{1}{2}\epsilon^4}{3\epsilon - 3\epsilon^2 + \epsilon^3} \right) = \frac{2}{3}.$$

A proper coarsening of the former, they would say, is $\frac{P(Red\ HQ)}{P(Red)} \geq 1 - \epsilon$. Then setting

$$R_3 = \{(x,y) \in R : \frac{y}{y+z} \geq 1 - \epsilon\} = \{(x,y) : y \geq (1-\epsilon)(1-x)\},$$

one finds that

$$Q_l(Blue) = \lim_{\epsilon \to 0} \left(\frac{\int_{R_3} x(6x)\, dx}{\int_{R_3} 6x\, dx} \right) = \lim_{\epsilon \to 0} \left(\frac{\frac{1}{2}\epsilon}{\epsilon} \right) = \frac{1}{2}.$$

Our rejection of this treatment is based on several considerations. First, **GH** is (*extremely*) unrealistic. For example, it implies that Judy thinks that HQ's credence in $Blue$ is over 390 times as likely to fall in the interval [.267, .268) as it is to fall in the interval [0,.001). Surely, though, it is far more likely to fall in the latter (as it will whenever HQ has conclusive or near-conclusive evidence favoring Red).

Second, it isn't obvious how one might generalize **GH** to situations in which Judy holds different priors. What, for example, if Judy's prior probability in $Blue$ is $\frac{2}{5}$? The uniform distribution over quadruples that G&H claim is natural isn't available when Judy's prior isn't already uniform over the finite partition in question.

[7] Letting $\epsilon \to 0$ would effectively fix Judy's posterior at the conditional expectation of $g(x,y)$ on the line segment $y = \frac{3}{4}(1-x)$, $0 \leq x \leq 1$, with respect to the σ-algebra generated by the triangles $\{a(1-x) \leq y \leq b(1-x) : 0 \leq x \leq 1\}$, $0 \leq a < b \leq 1$. The inconsistency encountered in the last section concerning how to condition on the message $P(Red\ 2nd) = 0$ and its various equivalent formulations would be, on this view, an artifact of the $(y = 1 - x)$ hypotenus's status as an atom of at least three different σ-algebras of potential interest. G&H take this hypotenus to be a null set, so Judy's different values for the conditional expectation of $g(x,y)$ on it with respect to the different algebras isn't (so probabilists assure us) any cause for concern. (For discussion of a famous case of this "paradox"–conditionalization on a great circle–see, e.g., Jaynes 2003.)

Third, even assuming that Judy's priors are fixed, it isn't clear from **GH** what would change in response to the information that HQ has more or less expected information exposure. To illustrate, suppose that there were a second duty officer known by Judy to be an expert relative to the first. What might be Judy's distribution over *his* credences? More generally, what if there were a nested chain of experts of some finite length? Which distribution Judy should deem appropriate for HQ's credence function depends on how much additional information she thinks he might have. If HQ's expected information exposure is small, the distribution ought to be tightly concentrated around her own prior $(J(Blue), J(Red\ HQ), J(Red\ 2nd)) = (\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$. If HQ's expected information exposure is large, the distribution ought to be tightly concentrated about the extreme points $(1, 0, 0)$, $(0, 1, 0)$ and $(0, 0, 1)$.

The purport of the above concerns is clear. Judy needs to employ a three parameter family of distributions. Two of the parameters would encode her prior $(x, y, 1 - x - y)$, while the third would encode HQ's expected additional information exposure. I.R. Goodman and Hung T. Nguyen (1999) suggested that the family of *Dirichlet* distributions (i.e. densities of the form $f(x, y, z) = kx^a y^b z^c$) can be adapted to this end. The Dirichlet distributions consistent with the $(\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$ prior, for example, have densities $f(x, y, z) = kx^{2t-1}y^{t-1}z^{t-1}$. (Here t is the "third parameter"; G&H employ the $t = 1$ instance upon elimination of one variable.) Note that as t increases, these distributions become more and more concentrated around Judy's prior. High values of t, therefore, should correspond to cases in which Judy does not think HQ has much more information than she has. As t decreases, on the other hand, these distributions become more and more concentrated around the extreme measures. Low values of t, therefore, should correspond to cases in which Judy thinks there is a very good chance that HQ has something close to complete information.

What's lacking, however, is an explanation of why Judy's distribution over HQ's credences would evolve from one Dirichlet distribution to another (or even from one convex combination of Dirichlet distributions to another) in response to finding out that HQ had potentially acquired more information. Such explanations are available for classical applications of Dirichlet distributions. (A famous case is the Laplace rule of succession, in which one estimates the unknown bias of a coin.) If one's prior distribution for the bias of a three-sided die with outcome space $\{X, Y, Z\}$ is given by the density function $h(x, y, z) = k_1 x^a y^b z^c$ and one observes X on a sample roll, one's posterior density will have the form $h(x, y, z) = k_2 x^{a+1} y^b z^c$ (which is again Dirichlet). Without such a diachronic tale, we see no reason to think that this family of distributions is appropriate to the current application.

6. Heuristics and a simulation

In the sense we are interested in, regression to the mean occurs when one learns that information exposure that might have eliminated the non-actual cells of a partition has failed to eliminate at least two cells. Specifically, it occurs because smaller cells are (in the generic case) more vulnerable to elimination, and so receive a greater boost when they survive. For a simple example, suppose A has B on the ropes, with 98% of the chips, at the final table of a poker tournament. Assuming they are equal

in ability, B then has a 2% chance of winning. If now you step away from the game for a few minutes and return to find that (and only that) neither A nor B has yet been eliminated, you will now assign B posterior win probability greater than 2%.

To see how regression arises in Judy's case, let E be the event "at least one region has been eliminated by HQ". According to the reasoning we have been employing, in the generic setting larger cells should be considered less vulnerable to elimination, conditional on their being non-actual.[8] Letting J be Judy's credence function, then,

$$J(E|Blue) > J(E|Red\ HQ) = J(E|Red\ 2nd).$$

Put another way,

$$J(E^c|Blue) < J(E^c|Red\ HQ) = J(E^c|Red\ 2nd) = J(E^c|Red).$$

Since $J(Blue) = \frac{1}{2} = J(Red)$, meanwhile, it follows that $J(Blue \cap E^c) < J(Red \cap E^c)$. Imagine now that the message from HQ contains only the information whether or not at least one region has been eliminated. Judy's expection in $Blue$, conditional on E^c (i.e. no region is eliminated) is then

$$E(Blue|E^c) = \frac{J(Blue \cap E^c)}{J(E^c)} = \frac{J(E^c \cap Blue)}{J(E^c \cap Blue) + J(E^c \cap Red)} < \frac{1}{2}.$$

If now we alter our favored $\mathbf{3_B}$ interpretation of the original problem, so that HQ suppresses the exact value of $P(Red\ HQ|Red)$ (noting only that it lies strictly between zero and one), then the information Judy receives is precisely that E^c is the case. It follows that she should adopt posterior in $Blue$ strictly less than $\frac{1}{2}$.

Any sufficiently realistic genericity assumption should, therefore, elicit regression. We don't know of any *entirely* realistic genericity assumption, but do think that since HQ's credences derive from conditionalization on unknown, unpredictable evidence, an at least interesting model for their evolution is the process that is most often used to represent the movement of particles, prices or probabilities subject to large numbers of unknown, unpredictable influences: *Brownian motion*. Since (not, we think, fortuitously) Judy's credences regress under this model, it's worth a look.

To start, we follow Grove and Halpern in identifying the space of possible HQ credence functions with the convex hull (a regular pyramid) of the vectors $(1,0,0,0)$, $(0,1,0,0)$, $(0,0,1,0)$ and $(0,0,0,1)$ in 4-space. We let $((b_h(t), b_s(t), r_h(t), r_s(t))$ be the barycentric coordinates of a 3-dimensional standard Brownian motion on this pyramid at time t, with initial data $(b_h(0), b_s(0), r_h(0), r_s(0)) = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$. When the motion reaches a side or edge the motion becomes 2-dimensional or 1-dimensional on that side or edge, until terminating at an extreme point.

[8]There is good reason to believe that more "story completions" exhibit this property than exhibit the reverse. Certainly larger regions are dramatically less likely to be eliminated when they are "larger by disjunction", since one must eliminate all disjuncts in order to eliminate a disjunction. A 2-dimensional Brownian motion model (we explore a 3-dimensional Brownian motion below) provides another instance; as shown in (Ferguson, unpublished ms.), the probability that such a motion originating within an equilateral triangle at barycentric coordinates $(\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$ exits the triangle out the farthest side is $\approx .1421$. If such a motion is used to model the movement of HQ's credences on $\{Blue, Red\ HQ, Red\ 2nd\}$, Judy will adopt posterior probability $\approx \frac{1}{1-.1421} \approx .5828 < \frac{2}{3}$ for $Blue$ conditional on a Red subregion being eliminated prior to elimination of $Blue$.

For each $t > 0$ one now obtains a solution (consistent with protocol $\mathbf{3_B}$, which we favor), namely

$$Q_t(Blue) = E\Big(b_h(t) + b_s(t)\Big|\frac{r_h(t)}{r_h(t)+r_s(t)} = \frac{3}{4} \wedge b_h(t) + b_s(t) > 0\Big).$$

This solution assumes that HQ transmits his message having observed the Brownian motion over the interval $[0, t]$. (The parameter t therefore encodes "how much information" HQ has.) In practice Judy won't know t precisely, but will subscribe to some continuous distribution over its possible values. For simplicity, we will take the distribution to be uniform on $[0, T]$ for some value of T.

We describe a numerical simulation of the above via an ersatz four player poker game. For each iteration, the players begin with equal balances $b_h = b_s = r_h = r_s = \$250$. For each hand, a small random stake S is computed (we used $S = \frac{1}{5}(1 + a - b)$, where a and b are independent and uniformly distributed on the unit interval, but not greater than the minimum player balance). Each of the n uneliminated players antes $\$S$, and a winner is determined by a virtual roll of a fair n-sided die.

For $j = 0, 1, \ldots, 499$, we put

$$R_j = \Big[\frac{1}{2} + \frac{j}{1000}, \frac{1}{2} + \frac{j+1}{1000}\Big) \cup \Big(\frac{1}{2} - \frac{j-1}{1000}, \frac{1}{2} - \frac{j}{1000}\Big],$$

and $R_{500} = \{0, 1\}$. To simulate a standard Brownian motion over $[0, T]$, one should choose N so that N hands at the given stake distribution corresponds (in variance) to the continuous process being simulated at time T. (We ran simulations for $N = 75,000$ and $N = 250,000$.) One then computes, over each hand in a large number of game iterations, sample averages, for each R_j, of $b_h + b_s$ conditional on $\frac{r_h}{r_h+r_s} \in R_j$.

Data sets for $N = 75,000$ and $N = 250,000$ are depicted in Figure 1; Infomin's recommendations and the $\frac{1}{2}$ solution are shown for comparison. At $P(Red\ HQ|Red) = \frac{3}{4}$, the two simulations indicate posteriors in $Blue$ of $\approx .527$ and $\approx .492$, respectively, while at $P(Red\ HQ|Red) = 1$ the posteriors are $\approx .664$ at $N = 75,000$ and $\approx .653$ at $N = 250,000$.[9] That the latter posteriors fall short of $\frac{2}{3}$ is an example of what we have been calling "regression", though use of this term at the middle ranges of the reported value $P(Red\ HQ|Red)$ seems somewhat more dubious (cf. footnote 5).

7. Conclusion

Reasons why there cannot be a canonical solution to the Judy Benjamin problem ought by now to be clear. Since regression kicks in to the degree that Judy is "surprised" by the non-elimination of the various regions, her limiting posterior in $Blue$ is sensitive to the prior likelihood she attaches to such eliminations. That is, Judy's posteriors are sensitive to how much evidence she expects HQ to have acquired–and it certainly seems clear that there can be no canonical answer to the question of how much evidence Judy should expect HQ to have acquired.

[9]The posteriors at $R_{499} = (0, .001] \cup [.999, 1)$ were substantially lower, $\approx .628$ and $\approx .597$ for $N = 75,000$ and $N = 250,000$ respectively.

Conditionalization on a finite partition, on the other hand, is a "canonical" method of updating, not because it is appropriate in all circumstances (or even "in the mean") where one learns which cell of a partition obtains, but because it is appropriate in a particular sort of case that one identifies as "standard"–namely the case in which the agent is told, always and only, which cell of a given partition obtains. Selection effects like we have been looking at don't apply in such a case, because there is no chance of receiving partial or incomplete information[10]; each cell is eliminated with certainty, conditional on its being non-actual.[11]

References

Douven, Igor and Jan-Willem Romeijn. 2011. A New Resolution of the Judy Benjamin Problem, *Mind* 120:637-670.

Ferguson, Tom. Gambler's Ruin in Three Dimensions. Manuscript. Available at https://www.math.ucla.edu/ tom/papers/unpublished/gamblersruin.pdf. Accessed February 13, 2018.

Goodman, I.R. and Hung T. Nguyen. 1999. Probability updating using second order probabilities and conditional event algebra, *Information Sciences* 121:295-347.

Grove, Adam J. and Joseph Y. Halpern. 1997. Probability: Conditioning vs. Cross-entropy, *Proceedings of the 13th Annual Conference on Uncertainty in Artificial Intelligence*, San Francisco. Morgan Kaufmann. pp. 208-214.

Jaynes, E.T. 2003. The Borel-Kolmogorov paradox. *Probability Theory: The Logic of Science.* Cambridge University Press. pp. 467-470.

Schervish, Mark J., Teddy Seidenfeld and Joseph B. Kadane. 2004. Stopping to Reflect, *Journal of Philosophy* 101:315-322.

Seidenfeld, Teddy. 1986. Entropy and Uncertainty. *Philosophy of Science* 53:467-491.

Van Fraassen, Bas C. 1981. A Problem for Relative Information Minimizers in Probability Kinematics. *The British Journal for the Philosophy of Science* 1981:375-379.

Van Fraassen, Bas C. 1984. Belief and the Will. *Journal of Philosophy* 81:235-256.

Department of Mathematical Sciences, University of Memphis
E-mail address: rmcctchn@memphis.edu

[10]It is natural to ask whether such effects arise in the incomplete information scenarios where *Jeffrey conditionalization* (as developed in R. Jeffrey 1965) is often taken to apply; if HQ reports that $P(Red\ 2nd) = \frac{1}{10}$, should Judy (as Jeffrey conditionalization suggests) adopt posterior credence in *Blue* equal to $\frac{3}{5}$, or to something else? If we treat the *Blue* subregions on par with the *Red* ones then the $\frac{3}{5}$ posterior is warranted by indifference. So such effects, if any, are sensitive to subdivision.

[11]Thanks to Michael Huemer and two anonymous referees for helpful suggestions, and also to the editors at *Synthese* for their persistence.

Figure 1: Brownian Motion Simulation