Category-based updating*

Jiaying Zhao
Princeton University

Daniel Osherson
Princeton University

October 31, 2012

Abstract

Given a prior distribution over a finite outcome space, how is the distribution updated when one outcome is excluded (i.e., assigned probability 0)? We describe two experiments in which estimated probabilities seem to "stick" to salient events. The probabilities of such events remain relatively invariant through updating. Our results suggest that the credence assigned to a salient category is sometimes more basic than the credence assigned to the constituents that comprise the category.

Introduction

Miss Marple witnessed the crime from afar but had the distinct impression that its author was a man. The only suspects were Albert, Bruce, Charles, David, Elizabeth, Florence, Gertrude, and Harriet, so Miss Marple gave higher probability to the men, resulting in the following distribution.

(1) Miss Marple's prior distribution:

Albert	Bruce	Charles	David	Elizabeth	Florence	Gertrude	Harriet
.17	.18	.17	.18	.07	.08	.07	.08

It subsequently emerged that David had a solid alibi, setting his probability to zero in Miss Marple's mind. To adjust the remaining probabilities, she reasoned as follows.

> "David is not the culprit but I'm just as convinced as before that the guilty party is a man. So I'll retain the probability of the *men* category by renormalizing the

*We thank Derek Shiller for a lecture on probability that inspired the work reported here. We also thank Vincenzo Crupi, Konstantinos Hadjichristidis, David Over, and Steven Sloman for helpful comments. Contact: Jiaying Zhao (jiayingz@princeton.edu).

probabilities of the three remaining men to add up to the same *men*'s probability as before. The probabilities of the women won't be altered."

Specifically, for each of Albert, Bruce, and Charles, Miss Marple multiplied the prior probability by

$$\frac{Pr\,(\text{Albert}) + Pr\,(\text{Bruce}) + Pr\,(\text{Charles}) + Pr\,(\text{David})}{Pr\,(\text{Albert}) + Pr\,(\text{Bruce}) + Pr\,(\text{Charles})}$$

to obtain:

(2) Miss Marple's posterior distribution:

Albert	Bruce	Charles	David	Elizabeth	Florence	Gertrude	Harriet
.2288	.2424	.2288	0	.07	.08	.07	.08

We see that the probabilities of *men* and *women* in (2) are .7 and .3, respectively, just as they were in (1).

Miss Marple's reasoning is compelling. David's exoneration does not alter her impression of having seen a man commit the crime; it just makes it more likely that one of the other men is guilty. Indeed, if Bruce and Charles came up with alibis as well, that would go to show that Albert is likely the culprit! It must be admitted, however, that a posterior distribution different from (2) results from thinking along Bayesian lines. A Bayesian would renormalize all seven remaining probabilities in light of David's zero. Specifically, each of the seven probabilities would be divided by their sum to reach:

(3) The Bayesian posterior distribution:

Albert	Bruce	Charles	David	Elizabeth	Florence	Gertrude	Harriet
.2073	.2194	.2073	0	.0854	.0976	.0854	.0976

The choice between (2) and (3) reflects different strategies for attaching credence to events. The Bayesian proceeds by distributing probabilities over outcomes in a sample space; an event inherits its probability from the sum of the probabilities of the outcomes that comprise it (see Hacking 2001). A different tradition sorts the available evidence by the specific event to which it is relevant, allowing an event to be directly supported even if its constituent outcomes are not (Shafer 1976). Miss Marple's reasoning seems more congenial to the latter approach inasmuch as she conserves her credence in *men* despite David's alibi undermining one of its elements.

It is not our present goal to judge Miss Marple.[1] Indeed, both her strategy and that of the Bayesian may prove useful in different contexts (Shafer and Tversky 1986). We wish merely to

[1]She is the amateur detective in Agatha Christie's crime novels.

identify some circumstances in which Marple-type reasoning (rather than Bayesian) seems to represent common intuition. For this purpose it will be helpful to introduce some conventions.

The experiments reported below elicit distributions over eight-member sample spaces. Each space decomposes into four-member salient categories analogous to *men* and *women* for Miss Marple. Typically, one of the categories attracts higher prior probability than its complement. In fact, we will only consider trials in which all the probabilities assigned to the high-probability category exceed all the probabilities assigned to the low-probability category. Such a prior distribution will be called *acceptable*, and is illustrated by (1). After an acceptable prior is established, we set one of its eight outcomes to zero probability, and invite the participant to update. The zero can be chosen from the high-probability or the low-probability category. In either case, *category-based updating* consists of replacing each of the three remaining probabilities in that category with the result of multiplying it by the ratio

$$\text{(4)} \qquad \frac{\text{sum of the four prior probabilities in the category}}{\text{sum of the three prior probabilities that were not set to zero in that category}}.$$

The probabilities in the complementary category are untouched. Category-based updating is illustrated in (2). By *Bayesian updating* we mean the usual operation, illustrated in (3).

For each acceptable prior distribution produced by an experimental participant, we ask whether the participant's posterior is numerically closer to the category-based versus the Bayesian update. It will be seen that category-based is closer but only if the outcome that is given zero probability is drawn from a high-probability category; otherwise, neither update approximates the participant's posterior better than the other.

Experiment 1

Method

Forty-one Princeton undergraduates participated in exchange for course credit. The experiment was built around the eight objects shown at the top of Table 1. Participants were first familiarized with the objects, then completed eight trials. Each trial had two parts, yielding prior and posterior distributions.

Creating a subjective prior distribution: For each trial, 100 objects were randomly drawn with replacement according to the trial's probability distribution, shown in the corresponding row of Table 1 (each row represents one trial). Each drawn object was presented for 100 ms at the center of a computer screen, followed by a 50 ms blank interval. The rapid sampling prevented counting the occurrences of the eight objects. After the 100 draws were completed, the participant was told that a further 101st random draw had been completed covertly. For each of the eight objects, the participant was asked to estimate its probability of being the 101st draw. These eight estimates (one for each object) constitute the *prior distribution* for the trial in question.

.20	.20	.20	.20	.05	.05	.05	.05	2
.05	.05	.05	.05	.20	.20	.20	.20	7
.20	.20	.20	.20	.05	.05	.05	.05	6
.05	.05	.05	.05	.20	.20	.20	.20	3
.20	.20	.05	.05	.20	.20	.05	.05	5
.05	.05	.20	.20	.05	.05	.20	.20	4
.20	.20	.05	.05	.20	.20	.05	.05	3
.05	.05	.20	.20	.05	.05	.20	.20	2

Table 1: **Stimuli for Experiment 1.** Each row corresponds to one trial, and shows the probability of a given object being selected in a single draw. (There were 100 draws per trial.) The rightmost column shows which object was revealed *not* to be the one selected in the covert 101st draw. (The objects are numbered from left to right.)

Creating a subjective posterior distribution: To complete the trial, the participant was next instructed that a certain object was *not* the one sampled in the 101st draw. The right hand column of Table 1 shows which object was excluded for each trial. The participant then re-estimated the probability of each of the seven objects being the 101st draw. Probability zero was filled in for the excluded object; the participant was asked to produce the remaining seven estimates. The resulting probabilities (including the zero) constitute the *posterior distribution*.

When estimating both the prior and posterior distributions, participants were allowed to proceed only if their estimates summed to unity. The eight trials were administered to participants in individually random order.

Results

In a given trial, a four-member subset of the eight objects is called *major* just in case:

(a) the members of the subset are either of the same color (all blue or all green), of the same shape (all triangles or all pentagons), or of the same line texture (solid or dotted);[2] and

(b) the prior probabilities assigned to members of the subset are each greater than all prior probabilities assigned to the remaining four objects.

[2]Major categories based on line texture were rare because the objective probabilities never favored such a grouping.

The complement of a major category is called *minor*. It is easy to verify that there is at most one way to divide the eight objects into major and minor categories. A trial will be called *acceptable* if it produced major/minor categories, and none of the prior probabilities were taken to be zero. All other trials were dropped from further analysis. An acceptable trial in which the excluded object (i.e., the one set to zero by the experimenter) falls in the major category will be called *a major trial*. The remaining trials (in which the excluded object falls in the minor category) are called *minor trials*. (In our story, Miss Marple is confronted with a major trial based on the category *man*.) The 41 participants produced 131 major trials and 124 minor trials.

Each acceptable trial is associated with a *Bayesian* and with a *category-based* update. The Bayesian update is the result of dividing each of the seven non-zeroed probabilities in the prior distribution by their sum. The category-based update is the result of multiplying each of the three non-zeroed prior probabilities from the category that holds the zeroed object by the ratio defined in (4); the probabilities of the category without the zeroed object are not changed. For each acceptable trial, the Bayesian and category-based updates were compared with the participant's posterior distribution. Specifically, we computed the mean absolute deviation between the eight numbers of the participant's posterior versus the Bayesian update; call this number the *Bayesian predictive error*. Likewise, we computed the mean absolute deviation between the eight numbers of the participant's posterior versus the category-based update; call this the *category-based predictive error*. We now present statistics about these two kinds of predictive errors. Major and minor trials are discussed separately.

Major trials. Of the 131 major trials produced in the experiment, 91 yielded posterior distributions with smaller category-based predictive error than Bayesian predictive error. The average category-based predictive error for major trials was 0.165 (SD = 0.12) whereas the average Bayesian predictive error was 0.181 (SD = 0.10). For each participant, we counted the number of major trials in which the category-based distribution had smaller predictive error than the Bayesian distribution. For 29 out of 41 participants, there were more major trials with smaller category-based predictive error than Bayesian predictive error. The reverse was true for seven participants, and five participants were tied (including one participant who produced no major trials). Finally, for each participant, we computed her average category-based predictive error as well as her average Bayesian predictive error on major trials. (The participant without major trials was excluded.) Across participants, the mean average category-based and Bayesian predictive errors were 0.167 (SD = 0.074) and 0.184 (SD = 0.065), respectively. They were reliably different via a paired t-test [$t(39) = 2.58$, $p = .01$].

Minor trials. Of the 124 minor trials produced in the experiment, 78 yielded posterior distributions with smaller Bayesian predictive error than category-based predictive error. The average category-based predictive error for minor trials was 0.094 (SD = 0.057) whereas the average Bayesian predictive error was 0.082 (SD = 0.041). For each participant, we counted the number of minor trials in which the category-based distribution had smaller predictive error than the Bayesian distribution. For 11 of the participants, there were more minor trials with

smaller category-based predictive error compared to Bayesian predictive error. The reverse was true for 23 participants, and seven participants were tied (including one participant who produced no minor trials). Finally, for each participant, we computed her average category-based predictive error as well as her average Bayesian predictive error on minor trials. (The participant without minor trials was excluded.) Across participants, the mean average category-based and Bayesian predictive errors were 0.094 (SD = 0.036) and 0.085 (SD = 0.031), respectively. A paired t-test produced a trend for smaller Bayesian predictive errors [$t(39) = 1.71$, $p = 0.10$].

Overall, Experiment 1 reveals a reliable tendency to preserve the probability of a major category when one of its members is excluded. In contrast, when a member of a minor category is excluded, the update shows a trend toward the Bayesian solution.

Experiment 2

In Experiment 1, the objective probabilities governing the sampling of objects were designed to encourage acceptable priors in the minds of participants. Experiment 2 relied on the participants' background knowledge for the same purpose.

Method

Thirty undergraduates (19 female, mean age 21.2 yrs, SD = 1.1) from Princeton University participated in exchange for course credit. Participants completed ten trials. Each trial had two parts, yielding prior and posterior distributions.

Creating a subjective prior distribution: In a given trial, eight familiar items were presented; the task was to assign each item its (subjective) probability of exceeding the other items along a certain criterion. For example, one trial presented four different headphones and four different GPS devices; for each item, participants gave their probability that it was the most expensive among the eight according to Amazon.com. Another trial presented four foreign (non-USA) cities and four USA cities; in this case, participants stated their probability that each city had the highest population among the eight. The ten trials are summarized in Table 2. The eight items of a given trial were presented simultaneously via computer monitor in individually randomized position.

Creating a subjective posterior distribution: In the second part of a trial the participant was informed that a certain item did not, in fact, exceed the others along the criterion for that trial. For example, they were informed that a particular GPS device was *not* the most expensive item among the eight indicated in the first row of Table 2. Participants then re-estimated the probabilities for the remaining seven items in the trial. These seven probabilities (plus zero for the excluded item) constitute the subjective posterior distribution. The last column in Table 2 shows which item was excluded in a given trial.

For both distributions, the participant was allowed to proceed only if her estimates summed

Items				Criterion	Excluded
headphone_1	headphone_2	headphone_3	headphone_4	highest price	6
GPS_1	GPS_2	GPS_3	GPS_4		
TV_1	TV_2	TV_3	TV_4	highest price	3
Air cond_1	Air cond_2	Air cond_3	Air cond_4		
budget car_1	budget car_2	budget car_3	budget car_4	listed price	7
luxury car_1	luxury car_2	luxury car_3	luxury car_4		
ivy-league_1	ivy-league_2	ivy-league_3	ivy-league_4	largest student body	8
state univ_1	state univ_2	state univ_3	state univ_4		
Seoul	Mumbai	Tokyo	Jakarta	greatest population	5
New York	Los Angeles	Chicago	Houston		
Malaysia	Singapore	S. Korea	China	greatest life expect.	2
Algeria	Turkey	Egypt	Morocco		
India	Indonesia	Taiwan	S. Korea	highest GDP	6
Sweden	Belgium	Switzerland	Netherlands		
Niger	Zambia	Uganda	Kenya	highest birth rate	6
Malaysia	Philippines	Vietnam	Fiji		
ExxonMobil	Apple	Chevron	Citigroup	highest 2011 profit	2
Shell	Gazprom	PetroChina	Samsung		
Avatar	Titanic	Star Wars	Dark Knight	largest USA box office	8
Toy Story	Finding Nemo	Lion King	Shrek		

Table 2: **Stimuli used in Experiment 2.** To illustrate, the second row corresponds to the trial in which four different brands of television and four different brands of air conditioners were displayed to participants via screen shots. For each item, participants gave their probability that it was the most expensive of the eight (according to Amazon.com). Subsequently, the participant was informed that the third item (TV_3) was not the most expensive, and they re-estimated probabilities for the remaining seven items. In a given trial, the authors' intuition points to a natural division between the two rows of four items.

to unity. The ten trials were administered to participants in individually random order.

Results

In each trial, the eight items may be intuitively divided into two sets of four. In trial 1, for example, the two sets are the four headphones versus the four GPS devices. Table 2 shows the divisions used. In a given trial, we qualify as *major* either of these intuitive subsets provided that each of the probabilities assigned to its members exceed each of the probabilities assigned to the members of its complement. The complement of a major subset is called *minor*. Similarly to Experiment 1, a trial will be called *acceptable* if it produced major/minor categories, and none of the prior probabilities were taken to be zero. All other trials were dropped from further

analysis. An acceptable trial in which the excluded item falls in the major category will be called *a major trial*. The remaining trials (in which the excluded item falls in the minor category) are called *minor trials*. The 30 participants produced 90 major trials and 60 minor trials.

Each acceptable trial is associated with a Bayesian and with a category-based update as explained above. For each acceptable trial, we computed the mean absolute deviation between the eight numbers of the participant's posterior versus the Bayesian update (the *Bayesian predictive error*). Likewise, we computed the mean absolute deviation between the eight numbers of the participant's posterior versus the category-based update (the *category-based predictive error*). We now present statistics about these two kinds of predictive errors for major and minor trials separately.

Major trials. Of the 90 major trials produced in the experiment, 64 yielded posterior distributions with smaller category-based predictive error than Bayesian predictive error. The average category-based predictive error for major trials was 0.135 (SD = 0.114) whereas the average Bayesian predictive error was 0.168 (SD = 0.080). For each participant, we counted the number of major trials in which the category-based distribution had smaller predictive error than the Bayesian distribution. For 20 out of 30 participants, there were more major trials with smaller category-based predictive error than Bayesian predictive error. The reverse was true for four participants, and six participants were tied. Finally, for each participant, we computed her average category-based predictive error as well as her average Bayesian predictive error on major trials. Across participants, the mean average category-based and Bayesian predictive errors were 0.133 (SD = 0.103) and 0.171 (SD = 0.068), respectively. They were reliably different via a paired t-test [$t(29) = 3.47$, $p = .002$].

Minor trials. Of the 60 minor trials produced in the experiment, 42 yielded posterior distributions with smaller Bayesian predictive error than category-based predictive error. The average category-based predictive error for minor trials was 0.119 (SD = 0.093) whereas the average Bayesian predictive error was 0.100 (SD = 0.068). For each participant, we counted the number of minor trials in which the category-based distribution had smaller predictive error than the Bayesian distribution. For seven of the participants, there were more minor trials with smaller category-based predictive error compared to Bayesian predictive error. The reverse was true for 17 participants, and six participants were tied (including two who produced no minor trials). Finally, for each of the 28 participants who produced minor trials, we computed her average category-based predictive error as well as her average Bayesian predictive error on minor trials. Across participants, the mean average category-based and Bayesian predictive errors were 0.119 (SD = 0.093) and 0.100 (SD = 0.068), respectively. A paired t-test produced a trend for smaller Bayesian predictive errors [$t(27) = 1.160$, $p = 0.256$].

Experiment 2 thus produced similar results to Experiment 1. Category-based updating prevailed when a member of a major category was excluded whereas there was a nonsignificant tendency to update in the Bayesian way for minor categories.

Discussion

Both experiments suggest that people often assign credence to complex events in a more fundamental way than they do to the event's atomic constituents. For, changes in the status of an event's constituent do not always propagate to the event itself. Thus, when the presentation of objects in Experiment 1 imparted high probability to the last draw being blue (for example), learning that a specific blue object was not drawn had relatively little impact on the belief in blue. Likewise, in Experiment 2, if a participant attached high probability to a European nation having highest GDP among the eight listed then they tended to retain that conviction even if one of the European nations was excluded as having highest GDP.

Such stability points to the use of "basic belief assignments" (Shafer, 1976; Smets and Kennes, 1994) when evaluating events defined over a finite sample space, rather than standard probability distributions. Given an event E made up of elementary outcomes $e_1 \dots e_n$, it may happen that belief in E is psychologically prior to belief in the e_i. In this case, credence flows "down" to the e_i from E, although there might also be a contribution from the e_i that prevents them from having uniform probability. This is different from the standard picture of credence flowing "up" to E from the conviction first garnered by each e_i. In a very helpful discussion, Kotzen (2012) puts the matter as follows (adapting to the present context). Credence flows down if the agent's *reason* for her belief in the e_i is her belief in E. Conversely, credence flows up if her reason for believing E is based on her beliefs about the e_i. Thus, in Experiment 1, the projection of many blue objects directly supported this category, and its probability flowed down to the more specific blue objects like the blue, solid triangle. The latter objects were apparently registered less directly. In Experiment 2, the GPS devices seemed generically more expensive than the headphones, and sent their credence down to specific GPS models.

On the other hand, our results suggest that low-probability events fail to provide reason to believe in their constituents. Indeed, minor categories gave no evidence of category-based updating, conforming instead to the Bayesian rule. The role of degree of belief in the choice of update might be clarified in future studies by posing direct questions about the probabilities of categories.

Consider again a major category C in our experiments. When the probability of an elementary event in C is set to zero, its prior probability p must be shifted to other events. We have focussed on just one alternative to Bayesian updating, namely, dividing p among the three remaining members of C in a proportional way, as described by (4). Alternatives come to mind, notably, dividing p into three equal parts and adding a part to each of the three remaining probabilities in C. Our neglect of the additive rule stems from its counter-intuitive behavior if one member of C has very low probability; updating additively can impart a posterior that is too high (Kotzen, 2012 offers a compelling example along with an additional objection to the additive rule). But it remains possible that some alternative to our version of category-based updating would improve the prediction of posterior probability.

References

Hacking, I. (2001). *An Introduction to Probability and Inductive Logic*. Cambridge University Press, Cambridge UK.

Kotzen, M. (2012). Where should displaced credence go? *unpublished*.

Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton University Press, Princeton NJ.

Shafer, G. and Tversky, A. (1986). Languages and designs for probability judgment. *Cognitive Science*, 9:309–339.

Smets, P. and Kennes, R. (1994). The transferable belief model. *Artificial Intelligence*, 66:191–243.