Safe Probability: restricted conditioning and extended marginalization

Peter Grünwald

CWI, Amsterdam and Leiden University, The Netherlands
pdg@cwi.nl

Abstract. Updating probabilities by conditioning can lead to bad predictions, unless one explicitly takes into account the mechanisms that determine (1) what is observed and (2) what has to be predicted. Analogous to the *observation*-CAR (coarsening at random) condition, used in existing analyses of (1), we propose a new *prediction task*-CAR condition to analyze (2). We redefine conditioning so that it remains valid if the mechanisms (1) and (2) are unknown. This will often update a singleton distribution to an imprecise set of probabilities, leading to dilation, but we show how to mitigate this problem by marginalization. We illustrate our notions using the Monty Hall Puzzle.

1 Introduction

Let P be a probability distribution on some space $\mathcal{Y}$. Suppose we are given information in the form of an event $B \subset \mathcal{Y}$. We are then asked to give the probability of another event $A \subset \mathcal{Y}$, given information B. Many people would be inclined to say "this probability is equal to $P(A|B)$, defined as $P(A,B)/P(B)$; this is just the standard definition of conditional probability". In this paper, we boldly propose a little extension of probability theory, in which we always have to make an additional calculation, to check whether predicting with $P(A|B)$ is *valid*, or at least *safe*. If it is unsafe, we should not use $P(A|B)$; we then risk getting answers that are wrong under any reasonable operational interpretation of probability. We explain this in Section 2, right after Example 3, and give formal definitions of safety and validity in Section 4.1, Definition 1; for now, we just note that unsafety implies there are other ways of updating P based on B that provably lead to better predictions. Indeed, we identify realistic situations in which updating by a "predictive distribution" $\tilde{P}$ *different* from standard conditioning is "safe", whereas standard conditioning is "unsafe".

All this may sound worrisome, especially to Bayesian readers: isn't there a plethora of evidence (by e.g. Savage (1954) and many, many others), axiomatic and otherwise, implying that conditioning is the *only* reasonable way to update probabilities? The answer is: yes, there is, and all our 'safe' updates are in fact compatible with conditioning if we were to work in a larger sample space $\mathcal{Z}$ that takes explicitly into account the *observation selection mechanism* (OSM) and the *task selection mechanism* (TSM). Here a 'task' can be any decision, prediction, or inference problem. Earlier work on OSM has been done within the CAR (coarsening at random) literature (Heitjan and Rubin, 1991, Grünwald and Halpern, 2003, De Cooman and Zaffalon, 2004). We generalize CAR-based OSM's and connect them to TSM's, which, to the best of our knowledge, have not been studied before. In practice such selection mechanisms, while relevant, may often be unknown, so we do not know the appropriate distribution P^* on $\mathcal{Z}$. We only know that P^* must be a member of some set of distributions $\mathcal{P}^*$, consisting of all distributions on $\mathcal{Z}$ that satisfy some known constraints. The 'safe' predictive distributions $\tilde{P}$ that we advocate typically coincide with a marginal distribution corresponding to some specific, special distribution in the set $\mathcal{P}^*$.

In the remainder of this introduction, we describe two well-known probability puzzles that motivate our research. Section 2 summarizes relevant insights from the CAR literature. Our original contributions are in Section 3 and beyond, in which we develop two notions of safety: the strong *guaranteed-validity* notion and a weaker notion which we just call *safety.* In the final section we return to the two puzzles to see what safe probability implies for them. Mathematical proofs and further discussion will be provided in the full paper of which this submission is an extended abstract.

Example 1. [**Monty Hall Puzzle**] (vos Savant, 1994, Gill, 2011) Suppose that you're on a game show and given a choice of three doors, named a, b and c. Behind one is a car; behind the others are goats. You pick door a. Before opening door a, Monty Hall, the quiz master (who knows what is behind each door) opens one of the other doors (say, door c), which has a goat. He then asks you if you still want to take what's behind door a, or to take what's behind the closed door (door b, in our case) instead. Should you switch? You may assume that, initially, the car was equally likely to be behind each of the doors, so it seems natural to define a sample space $\mathcal{Y} = \{a, b, c\}$ where $Y = y$ indicates that the car is behind door a, and $P(a) = P(b) = P(c) = 1/3$. You observe that the car is not behind door c, i.e. the remaining possibilities are $\{a, b\}$, Conditioning now gives that $P(b \mid \{a, b\}) = (1/3)/(2/3) = 1/2$, which suggests that the car is now equally likely to be behind door a and door b. Thus, there seems no reason to switch.

Now, 23 years after this problem was popularized, almost everybody agrees that this simple answer is wrong: as vos Savant pointed out, it is strongly in your interest to switch. However, initially, most people who heard about the puzzle, including some professors of probability theory (see (vos Savant, 1994)), were very hard to convince of this. It is here that safe probability can be useful: from the definition of safety in Section 3, one *immediately* sees that conditioning as we did above is 'unsafe', implying it will lead to suboptimal decisions. Briefly, for general spaces $\mathcal{Y}$, if the set of events $\mathcal{X}$ on which you can condition is not a partition of $\mathcal{Y}$, then conditioning on any of these events is unsafe. In the present case, the set of events is $\mathcal{X} = \{\{a, b\}, \{a, c\}\}$ (the latter would be observed if the quiz master had opened door b). The two events overlap (a is a member of both), hence do not form a partition, and hence you should not update by conditioning. This part is only of limited novelty — it has been argued before by many authors (perhaps most notably Shafer (1985)) that updating by conditioning only makes sense if a protocol is specified (corresponding to what we call an 'observation selection mechanism' below). Shafer (1996) formalizes this in terms of event trees which implicitly require conditioning events to form a partition. The only novelty here is our insight that, in practical cases in which the 'correct' event tree may be hard to construct, checking for overlap provides a *very simple sanity check* which *immediately* indicates that a problem is represented in a space in which conditioning makes no sense.

The real novelty of safe probability relates to the question whether, if the quiz master opens door c, the probability that the car is behind a *remains* 1/3. If one assumes that, in those cases in which the car is actually behind door a, the quiz master tosses a fair coin to decide whether to open door b or c, then the answer that $P(a)$ remains 1/3 is valid. However, it is unclear whether in the game as it was actually played on TV, Monty Hall really tossed a fair coin; for all we know he might have followed a very different rule, for example, open door c whenever you can. Previous analyses such as by Grünwald and Halpern (2003) that take into account that Monty's protocol is unknown, conclude that after the quiz master opened door b or c, a precise probability of the car being behind door a cannot be given any more: it can be anything between 0 and 1/2. In other words, the probability has *dilated* (Seidenfeld and Wasserman, 1993): it seems that, by observing additional information, one knows less than before (this will be explained in Example 2). But this does not seem satisfactory either: many people would reason that, since the quiz master in fact *has to* open door b or c, he gives no information about a, so the probability should remain 1/3. Using safe probability we can partially vindicate this intuition: we show that 1/3 does have a special status, even if the quiz master's protocol is unknown — the reasoning is, to some extent, correct after all, if our goal is just to asses whether the car is behind door a: let $Y' = 1$ iff a obtains, $Y' = 0$ otherwise. In Section 4 we show that $\tilde{P}$ defined by $\tilde{P}(Y' = 1 \mid \{a, b\}) = \tilde{P}(Y' = 1 \mid \{a, c\}) = 1/3$ is a sort of *marginal* distribution, and we show that predictions based on $\tilde{P}$ will behave exactly as they would if $\tilde{P}$ were actually the correct conditional distribution. Hence it is *safe* to act as if the probability remains 1/3.

2 The Problem with Overlapping Sets

Notation All sets we introduce below are finite. All probability distributions mentioned below are defined on $\mathcal{Z}$, our generic symbol for the sample space. A random variable (RV) is any function from $\mathcal{Z}$ to some arbitrary finite set. For a given RV we denote its range in calligraphic script. For example, a RV X maps $z \in \mathcal{Z}$ to $\mathcal{X}$. For RV Y with range $\{y_1, \ldots, y_m\}$, when we write $P(Y)$ we really mean the vector $(P(y_1), \ldots, P(y_m))$, where $P(y)$ abbreviates $P(Y = y)$.

Example 2. [**Dice**] This example is really just Monty Hall, without any misleading aspects. Suppose you and me play the following game: I roll a die, which we both know to be fair, i.e. $\mathcal{Z} = \mathcal{Y} = \{1, \ldots, 6\}$. I get to see the outcome, but you don't. I only tell you whether the outcome is below 3 or not, i.e. whether $Y \in \{1, 2\}$ or $Y \in \{3, 4, 5, 6\}$. Given this information, you are asked to give the probability that $Y = 3$. We agreed beforehand that, after throwing the die, I will tell you exactly one of the two statements, and that I won't lie. If I tell you $\{1, 2\}$, you would probably answer 'the probability of 3 is now 0', and if I tell you $\{3, 4, 5, 6\}$, you would say 'the probability of 3 is now 1/4'. This is the answer you get by conditioning: $P(Y = 3 \mid Y \in \{1, 2\}) = 0$ and $P(Y = 3 \mid Y \in \{3, 4, 5, 6\}) = 1/4$, and here it is obviously valid.

But now let's slightly change the game: we now agree beforehand that, after throwing the die, I will tell you either "$Y \in \{1, 2, 3, 4\}$" or "$Y \in \{3, 4, 5, 6\}$". Suppose that, when we actually play, I tell you $Y \in \{3, 4, 5, 6\}$. Given this observation, what is now the probability of 3? Many people would still say 1/4 but this answer is wrong. To see this, note that if, after throwing the die, I observe outcome 3 or 4, then I have a *choice* in what to tell you, and you do not know how I choose. For example, I may decide to always say $\{1, 2, 3, 4\}$ whenever I observe 3 or 4. In that case, if I say $Y \in \{3, 4, 5, 6\}$, the actual probability that $Y = 3$ is 0 rather than 1/4! (for if I had observed 3, I had certainly told you $\{1, 2, 3, 4\}$). Even if I am 'fair', i.e., when I observe 3 or 4, I flip a fair coin to decide whether to tell you $\{1, 2, 3, 4\}$ or $\{3, 4, 5, 6\}$, the answer 1/4 is still wrong: as we calculate below in (2), the probability of $Y = 3$ given $\{3, 4, 5, 6\}$ then becomes 1/6. Note that when we write '1/4 is invalid' we do not refer to the mathematical definition of conditional probability (the statement $P(Y = 6 \mid \{3, 4, 5, 6\}) = 1/4$ is after all a correct application of the definition of conditional probability). We explain what 'invalid' means here right after Example 3.

As explained by e.g. Grünwald and Halpern (2003) (GH from now on), to formalize problems such as this correctly, we need to move to a larger sample space in which we can explicitly represent the fact that I sometimes have a choice in what to tell you. This can be done by representing the problem in the space $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$, where $\mathcal{Y}$ is the *outcome space* as before, and $\mathcal{X}$ is the *observation* space, with associated RVs Y and X, respectively. $\mathcal{Z}$ was called the "sophisticated space" by GH. We assume (uncontroversially, see e.g. (Heitjan and Rubin, 1991)) that in this larger space, conditioning is the valid thing to do. In our case, $\mathcal{Y} = \{1, \ldots, 6\}$ as before, and $\mathcal{X} = \{\{1, 2, 3, 4\}, \{3, 4, 5, 6\}\}$. We know that the distribution P on $\mathcal{Z}$ must be compatible with the distribution on $\mathcal{Y}$, and we also agreed that I don't lie, so in our case this means that $P(Y = y) = 1/6$ for all $y \in \mathcal{Y}$, and $P(Y \in x \mid X = x) = 1$ for both $x \in \mathcal{X}$. This is not sufficient to specify $P((x, y))$ for all $(x, y) \in \mathcal{Z}$. For this, we would need two more probabilities p and q in $[0, 1]$, defined by setting

$$P(X = \{3, 4, 5, 6\} \mid Y = 3) = p, \; P(X = \{3, 4, 5, 6\} \mid Y = 4) = q. \tag{1}$$

Once we specify p and q, we can determine $P(x, y)$, and, more importantly for us, $P(y \mid x)$, for each $(x, y) \in \mathcal{Z}$.The interpretation is that when e.g. $Y = 3$, I flip a coin with bias p. If it lands heads I say $\{3, 4, 5, 6\}$, otherwise I say $\{1, 2, 3, 4\}$.

Example 3. We can now calculate the actual probability that $Y = 6$ given that I say $\{3, 4, 5, 6\}$ as

$$\begin{aligned} P(Y=6 \mid X=\{3,4,5,6\}) &= \tfrac{P(Y=6,X=\{3,4,5,6\})}{P(X=\{3,4,5,6\})} \\ &= \tfrac{P(6)}{P(3,X=\{3,4,5,6\})+P(4,X=\{3,4,5,6\})+P(5)+P(6)} \\ = \tfrac{P(6)}{P(3)P(X=3..6|3)+P(4)P(X=3..6|4)+P(5)+P(6)} &= \tfrac{\frac{1}{6}}{\frac{1}{6}\cdot(p+q+2)} = \tfrac{1}{p+q+2}, \end{aligned} \tag{2}$$

where in the third line we abbreviated all occcurrences of $Y = y$ to y, for $y \in \{3, \ldots, 6\}$. Suppose that we make no assumptions on p and q. This includes the deterministic cases (if $p = q = 1$ or $p = q = 0$) in which, when I have a choice, I'll *always* say the same thing. By varying p and q in (2), we find that $P(Y = 6 \mid X = \{3, 4, 5, 6\})$ can take on any value between 1/4 and 1/2, depending on the value of p and q.

All this shows that conditioning cannot always be valid. The meaning of 'valid' can be understood in three ways: (I) frequentist: conditioning is not *calibrated* , i.e. if we were to repeat the game of Example 2 independently many times, each time casting the die anew, then conditional relative frequencies will not converge to the corresponding conditional probabilities. For example, if I follow the strategy with $p = q = 0$, and we play, say, 6000 times, then each time I say $\{3, 4, 5, 6\}$, you will say that the probability of 3 is now 1/4; but of all the (approximately 2000) times that I will say $\{3, 4, 5, 6\}$, the actual outcome will be 5 or 6, so the conditional frequency of 3 is 0 rather than 1/4. (II) (perhaps more appealing to a Bayesian): decision-theoretic: not surprisingly, in the light of (I), using the conditional distributions to make predictions about Y can be suboptimal; we will see several examples of this in the next sections. (III) As we just saw, even if we do assume that conditioning is valid if the problem is modelled in the large space $\mathcal{X} \times \mathcal{Y}$, which takes into account the protocol, then, even if the protocol is 'fair', conditioning in the small space, omitting the protocol, can be invalid.

The original space $\mathcal{Y}$ was called the 'naive space' by GH. We may now ask when conditioning in the naive space is valid. The answer is given by the *coarsening at random (CAR) condition* (Heitjan and Rubin, 1991). For us, only a partial characterization is important: let $\mathcal{X}$ be a collection of subsets of $\mathcal{Y}$ and P^* be a distribution on $\mathcal{Y}$, and suppose that there is 'no lying'. If $\mathcal{X}$ partitions $\mathcal{Y}$, then naive conditioning is valid, i.e. conditioning in the naive space must coincide with conditioning in the sophisticated space. If $\mathcal{X}$ does not partition $\mathcal{Y}$, then naive condititioning can always be invalid: there exist distributions P on $\mathcal{X} \times \mathcal{Y}$ with marginal on Y equal to P^* and $x \in \mathcal{X}$ such that $P^*(X = x) > 0$, $P^*(Y \mid Y \in x) \neq P(Y \mid X = x)$; see Prop. 4.1 and Theorem 4.4(b) of GH.

3 Towards Safe Probability

First Attempt to restrict Conditioning The partition result above suggests a very simple definition of 'validity': we say conditional probability $P(A \mid B)$ is undefined unless a set $\mathcal{B}$ with $B \in \mathcal{B}$ is specified; we then write $P(A \mid B)$ as

$P_{\mathcal{B}}(A \mid B)$. $\mathcal{B}$ is the set of alternative events that might have been observed instead of B. We could then simply define conditioning $P_{\mathcal{B}}(A \mid B)$ to be 'valid' iff $\mathcal{B}$ is a partition, and restrict conditioning to valid cases: if $\mathcal{B}$ is not a partition, then it is undefined. This would already take care of the sanity check for the Monty Hall problem (Example 1), but it falls short of dealing with the second issue in Example 1 (how to assess the probability $\tilde{P}(Y' \mid \{a,b\})$), as well as the more general type of prediction task selection problems we will encounter below. We found these issues a lot more amenable to a random-variable based treatment, so that is the direction we take below.

Random Variables and Partitions The main advantage of a RV treatment is that the problem of invalid conditioning goes away — to some extent — automatically, since for every arbitrary random variables X, there is a partition Π such that conditioning on the value of X is equivalent to conditioning on the element of the partition that obtains (trivial proof provided in full paper). Thus, by our preliminary definition of validity based on event-conditioning as above, conditioning on a fixed RV X must *always* be valid. Thus, we could *define* conditional probability as $P(Y \mid X)$ for fixed RVs Y and X, and leave probabilities of the sort P(event A $\mid$ event B) undefined. One might argue that under such a definition of conditional probability, our problem of invalid conditioning goes away automatically. But it is more complicated than that: the problem goes away automatically *only* if it is implicitly understood that the distribution P for which $P(Y \mid X)$ is specified, will *only* be used to make predictions or decisions about Y given the value of X, irrespective of the value of X that is actually observed. Thus, for example, if $Y = (Y_1, Y_2)$, it is not valid in general to make a prediction about just Y_1 if $X = a$ is observed, and a prediction about just Y_2 if $X = b$ is observed (see Example 5 below). Yet if the prediction problem at hand satisfies the implicit *fixed* X*, fixed* Y—requirement, then conditioning on a fixed RV is indeed valid. This requirement often holds in signal processing, information-theoretic and machine learning applications such as classification and regression with i.i.d. random design.

Beyond Fixed RVs However, in many other standard applications of probability, we routinely make predictions about *various* RVs $Y_1, Y_2, \ldots$ conditioned on *various* RVs $X_1, X_2, \ldots$, and it is not precisely specified on what grounds a specific X_s or Y_t is chosen. For example, the Monty Hall and dice example can be interpreted in this way, as we show in Example 7. As a practically more relevant example, Bayes nets are often used to compute, e.g., how the probability that a patient has a certain disease would change counterfactually if (a) we were to observe that $X_1 = x_1$, or (b) we were to observe that $X_2 = x_2$; the result is then used to determine whether we should, in fact, observe RV X_1 or RV X_2 — both X_1 and X_2 may correspond to costly medical tests, and we may want to avoid doing two tests rather than one.

We only get away with such applications of probability if particular additional independence assumptions hold, which are usually left implicit. Rather than relying on such tacitly made assumptions to hold, as is usually done, it seems safer to use probability in a way which forces us to explicitly represent the *task selection mechanism* TSM (which determines what Y_j is observed) and the *observation selection mechanism* OSM (which determines what X_i is observed), so that we cannot violate our assumptions by mistake (which happens in the invalid $P(b \mid \{a,b\})$ answer to the Monty Hall problem, Ex. 1) or unnecessarily dilate a distribution (as happens in the Monty Hall problem when the probability $P(Y' = 1 \mid \{a,b\})$ is merely assessed to be in $[0, 1/2]$ instead of $1/3$). We now develop such an explicit representation of OSMs and TSMs.

4 Observation and Task Selection Mechanisms

We start with two examples that motivate the general definitions further below. Example 4 concerns OSMs: we show that event-based conditioning with overlap in the conditioning events (as in our three examples) can be rephrased as conditioning on a RV X_S selected from a set of RV $\{X_s \mid s \in \mathcal{S}\}$, where S is itself random. S then represents the OSM. Example 5 then concerns TSMs that determine what random variable Y_T has to be predicted.

Additional Notation in This Section For an event $\mathcal{E} \subset \mathcal{Z}$, we define the *indicator random variable* $I_{\mathcal{E}}$ to be 1 if $\mathcal{E}$ holds and 0 otherwise. For distribution P on $\mathcal{Z}$ and RV U we define $\textsc{support}_P(U) = \{u \in \mathcal{U} : P(U = u) > 0\}$. For a set of distributions $\mathcal{P}^*$ on $\mathcal{Z}$ and RVs U, V, W on $\mathcal{Z}$ we write $U \perp_{\mathcal{P}^*} V \mid W$ iff U and V are *conditionally independent* given W, that is, if for all $P \in \mathcal{P}^*$, for all $(u, v, w) \in \textsc{support}_P(U, V, W)$, it holds $P(U = u \mid V = v, W = w) = P(U = u \mid W = w)$. We say that $P, P' \in \mathcal{P}^*$ *agree* on an event $\mathcal{E}$ if $P(\mathcal{E}) = P'(\mathcal{E})$. We write $\mathcal{P}^*(\mathcal{E})$ to denote the set $\{P(\mathcal{E}) : P \in \mathcal{P}^*\}$. If all $P \in \mathcal{P}^*$ agree on $\mathcal{E}$, the probability of $\mathcal{E}$ is known relative to $\mathcal{P}^*$ and we write $P^*(\mathcal{E})$ rather than $\mathcal{P}^*(\mathcal{E})$. For two RVs U, V on $\mathcal{Z}$, we write $U \rightsquigarrow V$ (“U determines V” or “U is a *coarsening* of V”) if there is a function f such that for all $z \in \mathcal{Z}$, $V(z) = f(U(z))$.

Example 4. [**Observation Selection**] We define $\mathcal{S} = \{a, b\}$ and set RV $X_a := \{1,2,3,4\}$ if $Y \in \{1,2,3,4\}$ and RV $X_a = \{5,6\}$ otherwise. We set $X_b := \{3,4,5,6\}$ if $Y \in \{3,4,5,6\}$ and $X_b = \{1,2\}$ otherwise. Example 2 is equivalent to a scenario in which you observe X_S, where S is set to a if $Y \in \{1,2\}$; S is set to b if $Y \in \{5,6\}$, and if $Y \in \{3,4\}$, then whether you observe X_a or X_b depends on my protocol. To this end, we define the extended sample space $\mathcal{Z} = \mathcal{Y} \times \mathcal{S}$. We then set $P^*(S = a \mid Y = 1) = P^*(S = a \mid Y = 2) = P^*(S = b \mid Y = 5) = P^*(S = b \mid Y = 6) = 1$, and $P^*(S = b \mid Y = 3) = p, P^*(S = b \mid Y = 4) = q$. S — which in this case is just my protocol — is an example of what we call an observation-selection mechanism. We set $\mathcal{P}^*$ to be the set of all distributions on $\mathcal{Z}$ of the form above. The fact that we now have a set, rather than a single distribution reflects our ignorance of the precise protocol. The resulting setting is equivalent to Example 3: for example, if $Y = 3$, we will, with probability $1 - p$, observe $\{1,2,3,4\}$. Note that $S = a$ iff RV X in Example 3 is equal to $\{1,2,3,4\}$, and $S = b$ iff $X = \{3,4,5,6\}$. Thus, observing S is equivalent to observing X and we see that $\mathcal{Z} = \mathcal{Y} \times \mathcal{S}$ as here has equivalent representative power as $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ as defined above Example 3.

Example 5. [**Task Selection**] Let $X \in \{0,1\}$ and $Y \in \{a,b,c\}$. Imagine your goal is to predict aspects of Y given X, where an 'aspect' is a function that is determined by Y — for example, you might want to either predict $Y_1 = I_{Y=a}$ or $Y_2 = I_{Y=b}$ — and 'predicting Y_j' means coming up with a distribution $\tilde{P}$ for Y_j ($\tilde{P}$ may then be used as the basis for making decisions about Y under various loss function in the standard way, i.e. you choose the act that minimizes expected loss under $\tilde{P}$). Some external process determines whether Y_1 or Y_2 should be predicted. This process is modelled by an additional RV $T \in \mathcal{T} = \{1,2\}$. T is what we call a (prediction) task selection mechanism. The idea is that, in any realization of the system, Y_T (rather than the full Y) has to be predicted. Suppose you represent your uncertainty on (X,Y,T) by a set of distributions $\mathcal{P}^*$, all of which agree on (X,Y); hence the marginal distribution $P^*(X,Y)$ is known but the dependencies between (X,Y) and T may not be known. For concreteness, let's take some $P^*(X,Y)$ such that $P^*(Y = a \mid X = 1) = 0.8, P^*(Y = b \mid X = 0) = 0.9$, $P^*(Y = a) = 0.6$. If you think that T is determined independently of Y (for example, I ask you to predict either Y_1 or Y_2, and you know that I make my choice on external grounds, without knowing X or Y or $P^*(X,Y)$ myself), then $\mathcal{P}^*$ would be the set of all distributions P on $\mathcal{Z}$ with $P(X,Y) = P^*(X,Y)$ and with $(X,Y) \perp_P T$. Yet, if you don't know how I determine what RV I ask you to predict, you may want to choose for $\mathcal{P}^*$ the set of *all* distributions P on (X,Y,T) with $P(X,Y) = P^*(X,Y)$.

In standard uses of probability, the process T is rarely modelled explicitly, and upon observing $X = x$ and being asked to predict Y_t, you may be tempted to predict Y_t with the conditional distribution $P^*(Y_t \mid X = x)$. But in fact, this standard procedure is only valid if you are indeed in the situation with $Y \perp_{\mathcal{P}^*} T$, i.e. the process determining what you are asked is independent of Y itself. For otherwise, it would, for example, be possible to ask you about Y_1 whenever $Y = a$ and to ask about Y_2 whenever $Y \neq a$. Then, when observing $X = 1$, you will predict Y_1 with distribution $P^*(Y_1 = 1 \mid X = 1) = 0.8$, whereas the probability of $Y_1 = 1$ given that you are asked about it is really 1. Clearly standard conditioning is once again invalid, unless some independences involving (T,X,Y) hold. It seems we implicitly must assume, when we condition, that 'something like' $T \perp_{\mathcal{P}^*} Y \mid X$ is the casee (this includes the case that T is constant, fixed in advance); see Definition 3 below for a sharper formulation. Indeed, consider a scenario B in which I always ask you to predict Y_1 whenever $X = 1$ and Y_2 whenever $X = 0$, i.e. $T = f(X)$ with $f(1) = 1$ and $f(0) = 2$. Then T still depends on (X,Y) but now $T \perp Y \mid X$ and indeed T can now be safely ignored: the answers $P(Y_1 = 1 \mid X = 1) = 0.8$ and $P(Y_2 = 1 \mid X = 0) = 0.9$ are now valid.

But now, suppose that X is hidden from you yet you are still asked to predict Y_t; I still play scenario B but you don't know this. It is then standard practice for you to use the *marginal* distribution of Y_t, $P^*(Y_t) := \sum_x P^*(Y_t, X = x)$. In this case, when you predict Y_1, you will say that $P(Y_1 = 1) = 0.6$ (the marginal) whereas in fact, because I asked you for Y_1, it is 0.8 in this case. The problem is that, since you don't condition on X, Y still depends on T and hence you cannot ignore T when predicting Y. Thus, standard marginalization can be invalid when T is not independent of Y — just as we saw that conditioning on X could be invalid when T is not conditionally independent of Y given X. Now a sufficient (not necessary, see Def. 3 below) condition for valid prediction is that $T \perp_{\mathcal{P}^*} Y$ (since we marginalize, there is no conditioning on X any more). *Whenever we marginalize a probability in a practical application, we implicitly make an assumption like this!*

4.1 Main Definitions and Main Result

As in the examples, in our definition of a *predictive system* below, we represent a situation by a set $\mathcal{P}^*$ rather than a single P^* to reflect our ignorance: we believe that one $P^* \in \mathcal{P}^*$ is true (in a more Bayesian interpretation, it is the appropriate representation of our uncertainty), but we do not know which one.

Definition 1. *Let $\mathcal{P}^*$ be a set of distributions on $\mathcal{Z}$, and let X, Y be RVs on $\mathcal{Z}$ with ranges $\mathcal{X}, \mathcal{Y}$ such that $(*)$ all $P^* \in \mathcal{P}^*$ agree on (X, Y). Let $\mathcal{S}, \mathcal{T}$ be finite sets, let $\{X_s \mid s \in \mathcal{S}\}$ be a collection of RVs on $\mathcal{Z}$ such that for all $s \in \mathcal{S}$, $X_s \rightsquigarrow X$; let $\{Y_t \mid t \in \mathcal{T}\}$ be a collection of RVs on $\mathcal{Z}$ such that for all $t \in \mathcal{T}$, $Y_t \rightsquigarrow Y$. We call the collection $\mathbf{PS} = (\mathcal{P}^*, \mathcal{Z}, \mathcal{S}, \mathcal{T}, \{X_s \mid s \in \mathcal{S}\}, \{Y_t \mid t \in \mathcal{T}\})$ a* predictive system*. We call any RV (typically denoted S) that maps $\mathcal{Z}$ to $\mathcal{S}$ an OSM for $\mathbf{PS}$; and any RV (denoted T) that maps $\mathcal{Z}$ to $\mathcal{T}$ a TSM for $\mathbf{PS}$.*

Thus, we consider a setting in which, by $(*)$, the distribution $P^*(X, Y)$ is known to the DM (decision-maker). Since (X, Y) determine all variables X_s that we may observe and all variables Y_t to be predicted, the distribution of these RVs is known as well. The DM observes $X_S = x$, i.e. $X_s = x$ is observed for some X_s; but the X_s whose value is presented, is itself determined, perhaps randomly, by OSM S. Given this observation, DM has to predict RV Y_T, i.e. specify a distribution on Y_t for a t which itself determined, perhaps randomly, by TSM T. *Since we specifically do not require that all $P^* \in \mathcal{P}^*$ agree on* (S, T), DM may be ignorant on the actual details of the distribution of (S, T). Our goal is to find out whether it makes sense for DM to predict Y_T given X_S, S, T based on distributions that ignore S and/or T — this is what actual DMs (people) usually do and we want to see when they can get away with it. In many cases S and/or T are not observed, so DM cannot even condition on them; also their distribution may be unknown (not all $P^* \in \mathcal{P}^*$ may agree on them), so in such cases DM cannot even marginalize them out; he can just ignore them by acting as if the randomly determined (S, T) are actually not random but fixed in advance. The standard predictive distribution used by such a DM upon observing x is thus given by

$$\tilde{P}_{\text{standard}}(y \mid x, s, t) := P^*(Y_t = y \mid X_s = x), \tag{3}$$

the conditional distribution of Y_T that would arise if T were fixed in advance to t and S were fixed in advance to s. Yet the 'correct' conditional distribution is a member of the set $\{P(Y_t = y \mid X_s = x, S = s, T = t) : P \in \mathcal{P}^*\}$, and $\tilde{P}_{\text{standard}}(y \mid x, s, t)$ may not be equal to it. We want to find out when it can be safely used any way — this is determined in Definition 2 and Theorem 1 below. Note that $\tilde{P}_{\text{standard}}$ can be calculated without knowing the *distribution* of T or S and in some cases even without knowing the realized *value* s (CAR settings, Example 6 below). Note also that $\tilde{P}_{\text{standard}}$ does not 'marginalize out' S or T; it just pretends they are not random at all.

As seen in Example 5, a DM sometimes likes to predict Y or Y_T based on the *marginal* distribution of Y, with X marginalized out. In our setting, if X is marginalized out and S and T are ignored, this marginal distribution becomes

$$\tilde{P}_{\text{marginal}}(y \mid x, s, t) := P^*(Y_t = y) = \mathbf{E}_{X_s \sim P^*}[P^*(Y_t = y \mid X_s)] \tag{4}$$

Note that this distribution can be calculated without knowing either x or s or the distribution of S or T, but the treatment is asymmetrical: S and T are just ignored, X is marginalized out. Below we will see that it is sometimes smart to use $\tilde{P}_{\text{marginal}}$ rather than $\tilde{P}_{\text{standard}}$ even in situations in which x is observable.

The following definition can be applied to more general predictive distributions $\tilde{P}_\phi$ defined as $\tilde{P}_\phi(y \mid x, s, t) := P^*(Y_t = y \mid \phi(X_s) = \phi(x))$, for some function $\phi : \bigcup_{s \in \mathcal{S}} \mathcal{X}_s \to \varPhi$ (the special case with $S \equiv T \equiv 0$ and $X_0 \equiv X$, $Y_0 \equiv Y$, so that the TSM and OSM play no role, corresponds to the notion of "$\mathcal{C}$-conditioning" from Grünwald and Halpern (2011) with $\phi(x) = \mathcal{C}(x)$). $\tilde{P}_{\text{marginal}}$ and $\tilde{P}_{\text{standard}}$ are the special cases that use $\phi(x) \equiv 1$ and $\phi(x) = x$ respectively; for overall notational consistency we always include argument x in $\tilde{P}_\phi(y \mid x, s, t)$, even for $\tilde{P}_{\text{marginal}}$ which doesn't really depend on x.

Definition 2. *We say that a predictive distribution $\tilde{P}$ is* guaranteed-to-be-valid *(GTBV) for $Y_T \mid X_S$ relative to a predictive system* $\mathbf{PS}$ *if for all $P \in \mathcal{P}^*$, all $s, t, x, y \in \text{SUPPORT}_P(S, T, X_s, Y)$,*

$$\tilde{P}(y \mid x, s, t) = P(Y_t = y \mid X_s = x, S = s, T = t). \tag{5}$$

We say that $\tilde{P}_\phi$ is safe *for $Y_T \mid X_S$ if for all $(s, t) \in \text{SUPPORT}_{\mathcal{P}^*}(S, T)$, for all $P \in \mathcal{P}^*$, all x, y with $(s, t, x, y) \in \text{SUPPORT}_P(S, T, X_s, Y)$,*

$$\tilde{P}_\phi(y \mid x, s, t) = P(Y_t = y \mid \phi(X_s) = \phi(x), S = s, T = t). \tag{6}$$

In the full paper we extend the definition of safety to general $\tilde{P}$, but below we only use it for $\tilde{P}$ equal to $\tilde{P}_\phi$ for some ϕ as above. Intuitively, when observing X_S and having to predict Y_T, we would ideally like to use a $\tilde{P}$ that is GTBV. However, when the distributions of S and/or T are unknown, we cannot always determine this $\tilde{P}$. In some cases, we may still have that $\tilde{P}_{\text{standard}}$ is GTBV; but this will in general only be the case if the OSM S and TSM T play no crucial role, as formalized in Theorem 1 below. If S cannot be ignored, then we cannot determine a GTBV $\tilde{P}$ any more; but, as also shown in Theorem 1, if S cannot but T *can* be ignored, we can resort to predicting by $\tilde{P}_{\text{marginal}}$ and our predictions will still be 'safe'. 'Safety' is the condition that we always implicitly have to assume any way whenever we want to use a marginal distribution. In a frequentist view, if we use a 'safe' $\tilde{P}_\phi$ to repeatedly predict Y_T given X_S, where the (X_S, Y_T) pairs are sampled i.i.d. from some $P^* \in \mathcal{P}^*$ (hence $\mathcal{P}^*$ is 'true'), then the data *will behave exactly as if $\tilde{P}_\phi$ were the true conditional distribution.* The only way to find out whether data behave differently than predicted by $\tilde{P}_\phi$ would be to test $\tilde{P}_\phi$ (use it to make predictions) in situations in which Y_t depends on (S, T) given $\phi(X_s)$, yet does not depend on (S, T) given X_s (as in the final example in Ex. 5, where $\phi(X_s) \equiv 0$). Yet, since for $\tilde{P}_\phi$ the left-hand-side in (6) is equal to $P^*(Y_t = y \mid \phi(X_s) = \phi(x))$, the definition of 'safe' rules out exactly such T. In a Bayesian interpretation, no Dutch book can be made against $\tilde{P}_\phi$ by an adversary, unless that adversary has information about X that gets lost under the coarsening $\phi(X)$.

Definition 3. *We say that T represents an* ignorable TSM *for $Y_T \mid X_S$ if for all $t \in \text{SUPPORT}_{\mathcal{P}^*}(T)$, $Y_t \perp_{\mathcal{P}^*} I_{T=t} \mid X_S$. We say that S represents an* ignorable OSM *for $Y_T \mid X_S$ if for all $s \in \text{SUPPORT}_{\mathcal{P}^*}(S)$, $Y_T \perp_{\mathcal{P}^*} I_{S=s} \mid X_s$.*

We encountered ignorable TSMs in Example 5, where we had $X_S \equiv X$ (so the OSM plays no role), and we suggested the simpler but unnecessarily strong condition $Y \perp_{\mathcal{P}^*} T \mid X$, which implies that for all supported t, $P^*(Y_t \mid T = t, X) = P^*(Y_t \mid X)$, which coincides with the form in Definition 3. The analogously defined ignorable OSMs are related to CAR (Example 6). In normal, day-to-day probability uses, if X is not observed, we would like to use the marginal distribution $\tilde{P}_{\text{marginal}} = \tilde{P}_\phi$ for $\phi \equiv 0$, but if the TSM is not ignorable for Y_T, i.e. for $Y_T \mid \phi(X)$, then the resulting predictions can be disastrous, as shown at the end of Ex. 5; marginalization if X is unobserved can only be justified if T is ignorable for $Y_T \mid \phi(X)$. Now we turn the argument on its head: if T is ignorable for $Y_T \mid \phi(X)$ and X *is* observed, but the set of conditional distributions $\mathcal{P}^*(Y \mid X)$ may lead to bad predictions because it is too widely dilated, then it is preferable to use $\tilde{P}_{\text{marginal}}$ — since it is safe and the testing process is ignorable, data will behave exactly as if $\tilde{P}_{\text{marginal}}$ were fully valid, as shown in Theorem 1, part 2:

Theorem 1. *Let* **PS** *be a predictive system. (1) Suppose that T is an ignorable TSM for $Y_T \mid X_S$ and that S is an ignorable OSM for $Y \mid X_S$. Then $\tilde{P}_{\text{standard}}$ is safe for $Y_T \mid X_S$. (2) Suppose that T is ignorable for $Y_T \mid \phi(X_S)$ for some function ϕ. Then (even if S is not ignorable for $Y \mid X_S$), $\tilde{P}_\phi$ is safe for $Y_T \mid X_S$.*

Example 6. [**Embedding Event-Based Conditioning**] We can extend the idea of Example 4 to represent general overlapping event-based conditioning scenarios to our predictive systems. Given any collection $\mathcal{X}$ of nonempty subsets of $\mathcal{Y}$, we may simply set $\mathcal{S} = \mathcal{X}$, set $\mathcal{Z} = \mathcal{Y} \times \mathcal{S}$ and define, for each $s \in \mathcal{S}$, the RV X_s by $X_s((y, s')) = s$ if $y \in s$ and $X_s((y, s')) = \mathcal{Y} \setminus s$ otherwise — thus $X_s = s$ iff $Y \in s$. Assuming a trivial task selection mechanism ($\mathcal{T} = \{1\}$, $Y_1 = Y$, only the fixed RV Y has to be predicted), this re-represents event-based conditioning in terms of RVs. Observing set y translates to observing $X_S = y$; $\tilde{P}_{\text{standard}}$ corresponds to naive conditioning, since now $\tilde{P}_{\text{standard}}(y \mid x, s, t) = P^*(Y_1 = y \mid X_s = x) = P^*(Y = y \mid X_s = s) = P^*(Y = y \mid y \in s)$. The CAR condition (end of Section 2) expresses under what conditions on $\mathcal{P}^*$ naive conditioning is valid, i.e., in our new language, when $\tilde{P}_{\text{standard}}$ coincides with the true conditional distribution $P^*(Y = y \mid X_S = x, S = s)$ (we assumed $T \equiv 1$ so T can be ignored). By Definition 2 and Theorem 1 we see that CAR is implied if S is an ignorable OSM. With a little more work one shows that, for every event-based conditioning problem, one can construct an S as above, leading to the conclusion that 'ignorable S' generalizes standard CAR; similarly, we can think of 'ignorable T' as a kind of general 'prediction-task CAR'.

Example 7. [**Conclusion: Monty Hall, revisited**] We can model Monty Hall as a predictive system as in Definition 1 in complete analogy to the dice example: $Y \in \{a, b, c\}$; we observe X_S, with $S \in \{1, 2\}$, $X_1 = \{a, b\}$ and $S = 1$ if door c is opene; and $X_2 = \{a, c\}$ and $S = 2$ otherwise. We set $T \equiv 1$, i.e. the prediction task is independent of (X, Y). We want to find the distribution of $Y_1 = I_{Y=a}$. Checking Def. 2 we find that $\tilde{P}_{\text{standard}}$ (naive conditioning, see above) is *unsafe* for $Y_1 \mid X_S$. Yet, by Theorem 1, $\tilde{P}_{\text{marginal}}$ is *safe* for $Y_1 \mid X_S$. Hence, $\tilde{P}_{\text{standard}}$ should be avoided; yet if the goal is to predict $Y_1 = I_{Y=a}$, we advocate the use of $\tilde{P}_{\text{marginal}}$: if all uncertainty can be represented by a single distribution P^* and X is observable, then it is always preferable to predict with $\tilde{P}_\phi$ with $\phi(X) \equiv X$ and not marginalize, since our predictions will be sharper. Yet if uncertainty is represented by a *set* $\mathcal{P}^*$, as here, then the set of true conditional distributions given X_S may be dilated; and then, as long as it is safe, updating by $\tilde{P}_\phi$ for coarser ϕ may be preferable. This is the case here, where $\mathcal{P}^*(Y_1 = 1 \mid \{a, b\}) = [0, 1/2]$ whereas $\tilde{P}_{\text{marginal}}(Y_1 = 1 \mid \{a, b\}) = 1/3$ is precise, undilated and safe — so let's use it!

But now let $Y_2 = I_{Y=b}$. Can we also say that $\tilde{P}(Y_2 \mid \{a, b\}) = 2/3$? It turns out that this is still 'safe', but in a weaker sense than before; this will be treated in the full paper.

Acknowledgements for the anonymous referees who gave insightful comments.

References

G. De Cooman and M. Zaffalon. Updating beliefs with incomplete observations. *Artificial Intelligence*, 159(1):75–125, 2004.

R. Gill. The three door problem...-s. Invited Contribution to Springer's International Encyclopeadia of Statistical Science, 2011.

P. D. Grünwald and J. Y. Halpern. Updating probabilities. *Journal of Artificial Intelligence Research (JAIR)*, 19:243–278, 2003.

P.D. Grünwald and J.Y. Halpern. Making decisions using sets of probabilities: Updating, time consistency, and calibration. *Journal of Artificial Intelligence Research (JAIR)*, 42:393–426, 2011.

D.F. Heitjan and D.B. Rubin. Ignorability and coarse data. *Annals of Statistics*, 19:2244–2253, 1991.

L.J. Savage. *The Foundations of Statistics.* Dover Publications, 1954.

T. Seidenfeld and L. Wasserman. Dilation for convex sets of probabilities. *The Annals of Statistics*, 21:1139–1154, 1993.

G. Shafer. Conditional probability. *International Statistical Review*, 53(3):261–277, 1985.

G. Shafer. *The art of causal conjecture.* The MIT Press, 1996.

M. vos Savant. *Ask Marilyn.* St. Martins Mass Market Paperback, 1994.