Journal of Artificial Intelligence Research 19 (2003) 243–278 Submitted 10/02; published 10/03

Updating Probabilities

Peter D. Grünwald PDG@CWI.NL
CWI (National Research Institute for Mathematics and Computer Science in the Netherlands)
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands
www.cwi.nl/~pdg

Joseph Y. Halpern HALPERN@CS.CORNELL.EDU
Computer Science Department
Cornell University, Ithaca, NY 14853, USA
www.cs.cornell.edu/home/halpern

Abstract

As examples such as the Monty Hall puzzle show, applying conditioning to update a probability distribution on a "naive space", which does not take into account the protocol used, can often lead to counterintuitive results. Here we examine why. A criterion known as CAR ("coarsening at random") in the statistical literature characterizes when "naive" conditioning in a naive space works. We show that the CAR condition holds rather infrequently, and we provide a procedural characterization of it, by giving a randomized algorithm that generates all and only distributions for which CAR holds. This substantially extends previous characterizations of CAR. We also consider more generalized notions of update such as Jeffrey conditioning and minimizing relative entropy (MRE). We give a generalization of the CAR condition that characterizes when Jeffrey conditioning leads to appropriate answers, and show that there exist some very simple settings in which MRE essentially never gives the right results. This generalizes and interconnects previous results obtained in the literature on CAR and MRE.

1. Introduction

Suppose an agent represents her uncertainty about a domain using a probability distribution. At some point, she receives some new information about the domain. How should she update her distribution in the light of this information? *Conditioning* is by far the most common method in case the information comes in the form of an event. However, there are numerous well-known examples showing that naive conditioning can lead to problems. We give just two of them here.

Example 1.1: The *Monty Hall puzzle* (Mosteller, 1965; vos Savant, 1990): Suppose that you're on a game show and given a choice of three doors. Behind one is a car; behind the others are goats. You pick door 1. Before opening door 1, Monty Hall, the host (who knows what is behind each door) opens door 3, which has a goat. He then asks you if you still want to take what's behind door 1, or to take what's behind door 2 instead. Should you switch? Assuming that, initially, the car was equally likely to be behind each of the doors, naive conditioning suggests that, given that it is not behind door 3, it is equally likely to be behind door 1 and door 2. Thus, there is no reason to switch. However, another argument suggests you should switch: if a goat is behind door 1 (which happens with probability 2/3), switching helps; if a car is behind door 1 (which happens with probability 1/3), switching hurts. Which argument is right? ∎

Example 1.2: The *three-prisoners puzzle* (Bar-Hillel & Falk, 1982; Gardner, 1961; Mosteller, 1965): Of three prisoners a, b, and c, two are to be executed, but a does not know which. Thus, a thinks that the probability that i will be executed is 2/3 for $i \in \{a, b, c\}$. He says to the jailer, "Since either b or c is certainly going to be executed, you will give me no information about my own chances if you give me the name of one man, either b or c, who is going to be executed." But then, no matter what the jailer says, naive conditioning leads a to believe that his chance of execution went down from 2/3 to 1/2. ∎

There are numerous other well-known examples where naive conditioning gives what seems to be an inappropriate answer, including the *two-children puzzle* (Gardner, 1982; vos Savant, 1996, 1994) and the *second-ace puzzle* (Freund, 1965; Shafer, 1985; Halpern & Tuttle, 1993).[1]

Why does naive conditioning give the wrong answer in such examples? As argued by Halpern and Tuttle (1993) and Shafer (1985), the real problem is that we are not conditioning in the right space. If we work in a larger "sophisticated" space, where we take the protocol used by Monty (in Example 1.1) and the jailer (in Example 1.2) into account, conditioning does deliver the right answer. Roughly speaking, the sophisticated space consists of all the possible sequences of events that could happen (for example, what Monty would say in each circumstance, or what the jailer would say in each circumstance), with their probability.[2] However, working in the sophisticated space has problems too. For one thing, it is not always clear what the relevant probabilities in the sophisticated space are. For example, what is the probability that the jailer says b if b and c are to be executed? Indeed, in some cases, it is not even clear what the elements of the larger space are. Moreover, even when the elements and the relevant probabilities are known, the size of the sophisticated space may become an issue, as the following example shows.

Example 1.3: Suppose that a world describes which of 100 people have a certain disease. A world can be characterized by a tuple of 100 0s and 1s, where the ith component is 1 iff individual i has the disease. There are 2^{100} possible worlds. Further suppose that the "agent" in question is a computer system. Initially, the agent has no information, and considers all 2^{100} worlds equally likely. The agent then receives information that is assumed to be true about which world is the actual world. This information comes in the form of statements like "individual i is sick or individual j is healthy" or "at least 7 people have the disease". Each such statement can be identified with a set of possible worlds. For example, the statement "at least 7 people have the disease" can be identified with the set of tuples with at least 7 1s. For simplicity, assume that the agent is given information saying "the actual world is in set U", for various sets U. Suppose at some point the agent has been

1. Both the Monty Hall puzzle and the two-children puzzle were discussed in *Ask Marilyn*, Marilyn vos Savant's weekly column in "Parade Magazine". Of all *Ask Marilyn* columns ever published, they reportedly (vos Savant, 1994) generated respectively the most and the second-most response.
2. The notions of "naive space" and "sophisticated space" will be formalized in Section 2. This introduction is meant only to give an intuitive feel for the issues.

told that the actual world is in $U_1, \ldots, U_n$. Then, after doing conditioning, the agent has a uniform probability on $U_1 \cap \ldots \cap U_n$.

But how does the agent keep track of the worlds it considers possible? It certainly will not explicitly list them; there are simply too many. One possibility is that it keeps track of what it has been told; the possible worlds are then the ones consistent with what it has been told. But this leads to two obvious problems: checking for consistency with what it has been told may be hard, and if it has been told n things for large n, remembering them all may be infeasible. In situations where these two problems arise, an agent may not be able to condition appropriately. ∎

Example 1.3 provides some motivation for working in the smaller, more naive space. Examples 1.1 and 1.2 show that this is not always appropriate. Thus, an obvious question is when it is appropriate. It turns out that this question is highly relevant in the statistical areas of *selectively reported data* and *missing data*. Originally studied within these contexts (Rubin, 1976; Dawid & Dickey, 1977), it was later found that it also plays a fundamental role in the statistical work on *survival analysis* (Kleinbaum, 1999). Building on previous approaches, Heitjan and Rubin (1991) presented a necessary and sufficient condition for when conditioning in the "naive space" is appropriate. Nowadays this so-called *CAR (Coarsening at Random)* condition is an established tool in survival analysis. (for overviews, see (Gill, van der Laan, & Robins, 1997; Nielsen, 1998).) We examine this criterion in our own, rather different context, and show that it applies rather rarely. Specifically, we show that there are realistic settings where the sample space is structured in such a way that it is impossible to satisfy CAR, and we provide a criterion to help determine whether or not this is the case. We also give a *procedural* characterization of the CAR condition, by giving a randomized algorithm that generates all and only distributions for which CAR holds, thereby solving an open problem posed by Gill et al. (1997).

We then show that the situation is worse if the information does not come in the form of an event. For that case, several generalizations of conditioning have been proposed. Perhaps the best known are *Jeffrey conditioning* (Jeffrey, 1968) (also known as *Jeffrey's rule*) and *Minimum Relative Entropy (MRE) Updating* (Kullback, 1959; Csiszár, 1975; Shore & Johnson, 1980) (also known as *cross-entropy*). Jeffrey conditioning is a generalization of ordinary conditioning; MRE updating is a generalization of Jeffrey conditioning.

We show that Jeffrey conditioning, when applicable, can be justified under an appropriate generalization of the CAR condition. Although it has been argued, using mostly axiomatic characterizations, that MRE updating (and hence also Jeffrey conditioning) is, when applicable, the *only* reasonable way to update probability (see, e.g., (Csiszár, 1991; Shore & Johnson, 1980)), it is well known that there are situations where applying MRE leads to paradoxical, highly counterintuitive results (Hunter, 1989; Seidenfeld, 1986; van Fraassen, 1981).

Example 1.4: Consider the *Judy Benjamin* problem (van Fraassen, 1981): Judy is lost in a region that is divided into two halves, Blue and Red territory, each of which is further divided into Headquarters Company area and Second Company area. A priori, Judy considers it equally likely that she is in any of these four quadrants. She contacts her own headquarters by radio, and is told "I can't be sure where you are. If you are in Red territory, the odds are 3:1 that you are in HQ Company area ..." At this point the radio gives out. MRE updating on this information leads to a distribution where the posterior probability of being in Blue territory is greater than 1/2. Indeed, if HQ had said "If you are in Red territory, the odds are $\alpha : 1$ that you are in HQ company area ...", then for all $\alpha \neq 1$, according to MRE updating, the posterior probability of being in Blue territory is always greater than 1/2. ∎

Grove and Halpern (1997) provide a "sophisticated space" where conditioning gives what is arguably the more intuitive answer in the Judy Benjamin problem, namely that if HQ sends a message of the form "if you are in Red territory, then the odds are $\alpha : 1$ that you are in HQ company area" then Judy's posterior probability of being in each of the two quadrants in Blue remains at 1/4. Seidenfeld (1986), strengthening results of Friedman and Shimony (1971), showed that there is *no* sophisticated space in which conditioning will give the same answer as MRE in this case. (See also (Dawid, 2001), for similar results along these lines.) We strengthen these results by showing that, even in a class of much simpler situations (where Jeffrey conditioning cannot be applied), using MRE in the naive space corresponds to conditioning in the sophisticated space in essentially only trivial cases. These results taken together show that *generally speaking, working with the naive space, while an attractive approach, is likely to give highly misleading answers.* That is the main message of this paper.

We remark that, although there are certain similarities, our results are quite different in spirit from the well-known results of Diaconis and Zabell (1986). They considered when a posterior probability could be viewed as the result of conditioning a prior probability on some larger space. By way of contrast, we have a fixed larger space in mind (the "sophisticated space"), and are interested in when conditioning in the naive space and the sophisticated space agree.

It is also worth stressing that the distinction between the naive and the sophisticated space is entirely unrelated to the philosophical view that one has of probability and how one should do probabilistic inference. For example, the probabilities in the Monty Hall puzzle can be viewed as the participant's subjective probabilities about the location of the car and about what Monty will say under what circumstances; alternatively, they can be viewed as "frequentist" probabilities, inferred from watching the Monty Hall show on television for many weeks and then setting the probabilities equal to observed frequencies. The problem we address occurs both from a frequentist and from a subjective stance.

The rest of this paper is organized as follows. In Section 2 we formalize the notion of naive and sophisticated spaces. In Section 3, we consider the case where the information comes in the form of an event. We describe the CAR condition and show that it is violated in a general setting of which the Monty Hall and three-prisoners puzzle are special cases. In Section 4 we give several characterizations of CAR. We supply conditions under which it is guaranteed to hold and guaranteed not to hold, and we give a randomized algorithm that generates all and only distributions for which CAR holds. In Section 5 we consider the case where the information is not in the form of an event. We first consider situations where Jeffrey conditioning can be applied. We show that Jeffrey conditioning in the naive space gives the appropriate answer iff a generalized CAR condition holds. We then show that, typically, applying MRE in the naive space does not give the appropriate answer. We conclude with some discussion of the implication of these results in Section 6.

2. Naive vs. Sophisticated Spaces

Our formal model is a special case of the multi-agent systems framework (Halpern & Fagin, 1989), which is essentially the same as that used by Friedman and Halpern (1997) to model belief revision. We assume that there is some external world in a set W, and an agent who makes observations or gets information about that world. We can describe the situation by a pair (w, l), where $w \in W$ is the actual world, and l is the agent's *local state*, which essentially characterizes her information. W is what we called the "naive space" in the introduction. For the purposes of this paper, we assume that l has the form $\langle o_1, \ldots, o_n \rangle$, where o_j is the observation that the agent makes at time j, $j = 1, \ldots, n$. This representation implicitly assumes that the agent remembers everything she has observed (since her local state encodes all the previous observations). Thus, we ignore memory issues here. We also ignore computational issues, just so as to be able to focus on when conditioning is appropriate.

A pair $(w, \langle o_1, \ldots, o_n \rangle)$ is called a run. A run may be viewed as a complete description of what happens over time in one possible execution of the system. For simplicity, in this paper, we assume that the state of the world does not change over time. The "sophisticated space" is the set of all possible runs.

In the Monty Hall puzzle, the naive space has three worlds, representing the three possible locations of the car. The sophisticated space describes what Monty would have said in all circumstances (i.e., Monty's *protocol*) as well as where the car is. The three-prisoners puzzle is treated in detail in Example 2.1 below. While in these cases the sophisticated space is still relatively simple, this is no longer the case for the Judy Benjamin puzzle. Although the naive space has only four elements, constructing the sophisticated space involves considering all the things that HQ could have said, which is far from clear, and the conditions under which HQ says any particular thing. Grove and Halpern (1997) discuss the difficulties in constructing such a sophisticated space.

In general, not only is it not clear what the sophisticated space is, but the need for a sophisticated space and the form it must take may become clear only after the fact. For example, in the Judy Benjamin problem, before contacting headquarters, Judy would almost certainly not have had a sophisticated space in mind (even assuming she was an expert in probability), and could not have known the form it would have to take until after hearing headquarter's response.

In any case, if the agent has a prior probability on the set $\mathcal{R}$ of possible runs in the sophisticated space, after hearing or observing $\langle o_1, \ldots, o_k \rangle$, she can condition, to get a posterior on $\mathcal{R}$. Formally, the agent is conditioning her prior on the set of runs where her local state at time k is $\langle o_1, \ldots, o_k \rangle$.

Clearly the agent's probability $\Pr$ on $\mathcal{R}$ induces a probability $\Pr_W$ on W by marginalization. We are interested in whether the agent can compute her posterior on W after observing $\langle o_1, \ldots, o_k \rangle$ in a relatively simple way, without having to work in the sophisticated space.

Example 2.1: Consider the three-prisoners puzzle in more detail. Here the naive space is $W = \{w_a, w_b, w_c\}$, where w_x is the world where x is not executed. We are only interested in runs of length 1, so $n = 1$. The set O of observations (what agent can be told) is $\{\{w_a, w_b\}, \{w_a, w_c\}\}$. Here "$\{w_a, w_b\}$" corresponds to the observation that either a or b will *not* be executed (i.e., the jailer saying "c will be executed"); similarly, $\{w_a, w_c\}$ corresponds to the jailer saying "b will be executed". The sophisticated space consists of the four runs

$$(w_a, \langle\{w_a, w_b\}\rangle), (w_a, \langle\{w_a, w_c\}\rangle), (w_b, \langle\{w_a, w_b\}\rangle), (w_c, \langle\{w_a, w_c\}\rangle).$$

Note that there is no run with observation $\langle\{w_b, w_c\}\rangle$, since the jailer will not tell a that he will be executed.

According to the story, the prior $\Pr_W$ in the naive space has $\Pr_W(w) = 1/3$ for $w \in W$. The full distribution Pr on the runs is not completely specified by the story. In particular, we are not told the probability with which the jailer will say b and c if a will not be executed. We return to this point in Example 3.2. ∎

3. The CAR Condition

A particularly simple setting is where the agent observes or learns that the external world is in some set $U \subseteq W$. For simplicity, we assume throughout this paper that the agent makes only one observation, and makes it at the first step of the run. Thus, the set O of possible observations consists of nonempty subsets of W. Thus, any run r can be written as $r = (w, \langle U\rangle)$ where w is the actual world and U is a nonempty subset of W. However, O does not necessarily consist of *all* the nonempty subsets of W. Some subsets may never be observed. For example, in Example 2.1, a is never told that he will be executed, so $\{w_b, w_c\}$ is not observed. We assume that the agent's observations are accurate, in that if the agent observes U in a run r, then the actual world in r is in U. That is, we assume that all runs are of the form $r = (w, \langle U\rangle)$ where $w \in U$. In Example 2.1, accuracy is enforced by the requirement that runs have the form $(w_x, \langle\{w_x, w_y\}\rangle)$.

The observation or information obtained does not have to be exactly of the form "the actual world is in U". It suffices that it is equivalent to such a statement. This is the case in both the Monty Hall puzzle and the three-prisoners puzzle. For example, in the three-prisoners puzzle, being told that b will be executed is essentially equivalent to observing $\{w_a, w_c\}$ (either a or c will not be executed).

In this setting, we can ask whether, after observing U, the agent can compute her posterior on W by conditioning on U. Roughly speaking, this amounts to asking whether observing U is the same as discovering that U is true. This may not be the case in general—observing or being told U may carry more information than just the fact that U is true. For example, if for some reason a knows that the jailer would never say c if he could help it (so that, in particular, if b and c will be executed, then he will definitely say b), then hearing c (i.e., observing $\{w_a, w_b\}$) tells a much more than the fact that the true world is one of w_a or w_b. It says that the true world must be w_b (for if the true world were w_a, the jailer would have said b).

In the remainder of this paper we assume that W is finite. For every scenario we consider we define a set of possible observations O, consisting of nonempty subsets of W. For given W and O, the set of runs $\mathcal{R}$ is then defined to be the set

$$\mathcal{R} = \{(w, \langle U\rangle) \mid U \in O, w \in U\}.$$

Given our assumptions that the state does not change over time and that the agent makes only one observation, the set $\mathcal{R}$ of runs can be viewed as a subset of $W \times O$. While just taking $\mathcal{R}$ to be a subset of $W \times O$ would slightly simplify the presentation here, in general, we certainly want to allow sequences of observations. (Consider, for example, an n-door version of the Monty Hall problem, where Monty opens a sequence of doors.) This framework extends naturally to that setting.

Whenever we speak of a distribution Pr on $\mathcal{R}$, we implicitly assume that the probability of any set on which we condition is strictly greater than 0. Let X_W and X_O be two random variables on $\mathcal{R}$, where X_W is the actual world and X_O is the observed event. Thus, for $r = (w, \langle U \rangle)$, $X_W(r) = w$ and $X_O(r) = U$. Given a distribution Pr on runs $\mathcal{R}$, we denote by $\Pr_W$ the marginal distribution of X_W, and by $\Pr_O$ the marginal distribution of X_O. For example, for $V, U \subseteq W$, $\Pr_W(V)$ is short for $\Pr(X_W \in V)$ and $\Pr_W(V \mid U)$ is short for $\Pr(X_W \in V \mid X_W \in U)$.

Let Pr be a prior on $\mathcal{R}$ and let $\Pr' = \Pr(\cdot \mid X_O = U)$ be the posterior after observing U. The main question we ask in this paper is under what conditions we have

$$\Pr'_W(V) = \Pr_W(V|U) \tag{1}$$

for all $V \subseteq W$. That is, we want to know under what conditions the posterior W induced by $\Pr'$ can be computed from the prior on W by conditioning on the observation. (Example 3.2 below gives a concrete case.) We stress that Pr and $\Pr'$ are distributions on $\mathcal{R}$, while $\Pr_W$ and $\Pr'_W$ are distributions on W (obtained by marginalization from Pr and $\Pr'$, respectively). Note that (1) is equivalently stated as

$$\Pr(X_W = w \mid X_O = U) = \Pr(X_W = w \mid X_W \in U) \text{ for all } w \in U. \tag{2}$$

(1) (equivalently, (2)) is called the "CAR condition". It can be characterized as follows[3]:

Theorem 3.1: *(Gill et al., 1997) Fix a probability* Pr *on* $\mathcal{R}$ *and a set* $U \subseteq W$. *The following are equivalent:*

(a) If $\Pr(X_O = U) > 0$, *then* $\Pr(X_W = w \mid X_O = U) = \Pr(X_W = w \mid X_W \in U)$ *for all* $w \in U$.

(b) The event $X_W = w$ *is independent of the event* $X_O = U$ *given* $X_W \in U$, *for all* $w \in U$.

(c) $\Pr(X_O = U \mid X_W = w) = \Pr(X_O = U \mid X_W \in U)$ *for all* $w \in U$ *such that* $\Pr(X_W = w) > 0$.

(d) $\Pr(X_O = U \mid X_W = w) = \Pr(X_O = U \mid X_W = w')$ *for all* $w, w' \in U$ *such that* $\Pr(X_W = w) > 0$ *and* $\Pr(X_W = w') > 0$.

For completeness (and because it is useful for our later Theorem 5.1), we provide a proof of Theorem 3.1 in the appendix.

The first condition in Theorem 3.1 is just (2). The third and fourth conditions justify the name "coarsening at random". Intuitively, first some world $w \in W$ is realized, and then

3. Just after this paper was accepted, we learned that there really exist two subtly different versions of the CAR condition: "weak" CAR and "strong" CAR. What is called CAR in this paper is really "weak" CAR. Gill, van der Laan, and Robins (1997) implicitly use the "strong" definition of CAR. Indeed, the statement and proof of Theorem 3.1 (about weak CAR) are very slightly different from the corresponding statement by Gill et al. (1997) (about strong CAR); the difference is explained in Section 6.

some "coarsening mechanism" decides which event $U \subseteq W$ such that $w \in U$ is revealed to the agent. The event U is called a "coarsening" of w. The third and fourth conditions effectively say that the probability that w is coarsened to U is the same for all $w \in U$. This means that the "coarsening mechanism" is such that the probability of observing U is not affected by the specific value of $w \in U$ that was realized.

In the remainder of this paper, when we say "Pr satisfies CAR", we mean that Pr satisfies condition (a) of Theorem 3.1 (or, equivalently, any of the other three conditions) *for all* $U \in O$. Thus, "Pr satisfies CAR" means that conditioning in the naive space W coincides with conditioning in the sophisticated space $\mathcal{R}$ with probability 1. The CAR condition explains why conditioning in the naive space is not appropriate in the Monty Hall puzzle or the three-prisoners puzzle. We consider the three-prisoners puzzle in detail; a similar analysis applies to Monty Hall.

Example 3.2: In the three-prisoners puzzle, what is a's prior distribution Pr on $\mathcal{R}$? In Example 2.1 we assumed that the marginal distribution $\Pr_W$ on W is uniform. Apart from this, Pr is unspecified. Now suppose that a observes $\{w_a, w_c\}$ ("the jailer says b"). Naive conditioning would lead a to adopt the distribution $\Pr_W(\cdot \mid \{w_a, w_c\})$. This distribution satisfies $\Pr_W(w_a \mid \{w_a, w_c\}) = 1/2$. Sophisticated conditioning leads a to adopt the distribution $\Pr' = \Pr(\cdot \mid X_O = \{w_a, w_c\})$. By part (d) of Theorem 3.1, naive conditioning is appropriate (i.e., $\Pr'_W = \Pr_W(\cdot \mid \{w_a, w_c\})$) only if the jailer is equally likely to say b in both worlds w_a and w_c. Since the jailer must say that b will be executed in world w_c, it follows that $\Pr(X_O = \{w_a, w_c\} \mid X_W = w_c) = 1$. Thus, conditioning is appropriate only if the jailer's protocol is such that he definitely says b in w_a, i.e., even if both b and c are executed. But if this is the case, when the jailer says c, conditioning $\Pr_W$ on $\{w_a, w_b\}$ is *not* appropriate, since then a knows that he will be executed. The world cannot be w_a, for then the jailer would have said b. Therefore, *no matter what the jailer's protocol is*, conditioning in the naive space cannot coincide with conditioning in the sophisticated space for both of his responses. ∎

The following example shows that in general, in settings of the type arising in the Monty Hall and the three-prisoners puzzle, the CAR condition can only be satisfied in very special cases:

Example 3.3: Suppose that $O = \{U_1, U_2\}$, and both U_1 and U_2 are observed with positive probability. (This is the case for both Monty Hall and the three-prisoners puzzle.) Then the CAR condition (Theorem 3.1(c)) cannot hold for both U_1 and U_2 unless $\Pr(X_W \in U_1 \cap U_2)$ is either 0 or 1. For suppose that $\Pr(X_O = U_1) > 0$, $\Pr(X_O = U_2) > 0$, and $0 < \Pr(X_W \in U_1 \cap U_2) < 1$. Without loss of generality, there is some $w_1 \in U_1 - U_2$ and $w_2 \in U_1 \cap U_2$ such that $\Pr(X_W = w_1) > 0$ and $\Pr(X_W = w_2) > 0$. Since observations are accurate, we must have $\Pr(X_O = U_1 \mid X_W = w_1) = 1$. If CAR holds for U_1, then we must have $\Pr(X_O = U_1 \mid X_W = w_2) = 1$. But then $\Pr(X_O = U_2 \mid X_W = w_2) = 0$. But since $\Pr(X_O = U_2) > 0$, it follows that there is some $w_3 \in U_2$ such that $\Pr(X_W = w_3) > 0$ and $\Pr(X_O = U_2 \mid X_W = w_3) > 0$. This contradicts the CAR condition. ∎

So when does CAR hold? The previous example exhibited a combination of O and W for which CAR can only be satisfied in "degenerate" cases. In the next section, we shall study this question for arbitrary combinations of O and W.

4. Characterizing CAR

In this section, we provide some characterizations of when the CAR condition holds, for finite O and W. Our results extend earlier results of Gill, van der Laan, and Robins (1997). We first exhibit a simple situation in which CAR is guaranteed to hold, and we show that this is the only situation in which it is guaranteed to hold. We then show that, for arbitrary O and W, we can construct a 0-1–valued matrix from which a strong necessary condition for CAR to hold can be obtained. It turns out that, in some cases of interest, CAR is (roughly speaking) guaranteed *not* to hold except in "degenerate" situations. Finally, we introduce a new "procedural" characterization of CAR: we provide a mechanism such that a distribution Pr can be thought of as arising from the mechanism if and only if Pr satisfies CAR.

4.1 When CAR is Guaranteed to Hold

We first consider the only situation where CAR is guaranteed to hold: if the sets in O are pairwise disjoint.

Proposition 4.1: *The CAR condition holds for all distributions* Pr *on $\mathcal{R}$ if and only if O consists of pairwise disjoint subsets of W.*

What happens if the sets in O are not pairwise disjoint? Are there still cases (combinations of O, W, and distributions on $\mathcal{R}$) when CAR holds? There are, but they are quite special.

4.2 When CAR May Hold

We now present a lemma that provides a new characterization of CAR in terms of a simple 0/1-matrix. The lemma allows us to determine for many combinations of O and W, whether a distribution on $\mathcal{R}$ exists that satisfies CAR and gives certain worlds positive probability.

Fix a set $\mathcal{R}$ of runs, whose worlds are in some finite set W and whose observations come from some finite set $O = \{U_1, \ldots, U_n\}$. We say that $A \subseteq W$ is an $\mathcal{R}$-*atom* relative to W and O if A has the form $V_1 \cap \ldots \cap V_n$, where each V_i is either U_i or $\overline{U}_i$, and $\{r \in \mathcal{R} : X_W(r) \in A\} \neq \emptyset$. Let $\mathcal{A} = \{A_1, \ldots, A_m\}$ be the set of $\mathcal{R}$-atoms relative to W and O. We can think of $\mathcal{A}$ as a partition of the worlds according to what can be observed. Two worlds w_1 and w_2 are in the same set $A_i \in \mathcal{A}$ if there are no observations that distinguish them; that is, there is no observation $U \in O$ such that $w_1 \in U$ and $w_2 \notin U$. Define the $m \times n$ matrix S with entries s_{ij} as follows:

$$s_{ij} = \begin{cases} 1 & \text{if } A_i \subseteq U_j \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

We call S the *CARacterizing matrix* (for O and W). Note that each row i in S corresponds to a unique atom in $\mathcal{A}$; we call this the atom *corresponding* to row i. This matrix (actually, its transpose) was first introduced (but for a different purpose) by Gill et al. (1997).

Example 4.2: Returning to Example 3.3, the CARacterizing matrix is given by

$$\begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \end{pmatrix},$$

where the columns correspond to U_1 and U_2 and the rows correspond to the three atoms $U_1 - U_2, U_1 \cap U_2$ and $U_2 - U_1$. For example, the fact that entry s_{31} of this matrix is 0 indicates that U_1 cannot be observed if the actual world w is in $U_2 - U_1$. ∎

In the following lemma, $\vec{\gamma}^T$ denotes the transpose of the (row) vector $\vec{\gamma}$, and $\vec{1}$ denotes the row vector consisting of all 1s.

Lemma 4.3: *Let $\mathcal{R}$ be the set of runs over observations O and worlds W, and let S be the CARacterizing matrix for O and W.*

(a) Let Pr *be any distribution over $\mathcal{R}$ and let S' be the matrix obtained by deleting from S all rows corresponding to an atom A with $\Pr(X_W \in A) = 0$. Define the vector $\vec{\gamma} = (\gamma_1, \ldots, \gamma_n)$ by setting $\gamma_j = \Pr(X_O = U_j \mid X_W \in U_j)$ if $\Pr(X_W \in U_j) > 0$, and $\gamma_j = 0$ otherwise, for $j = 1 \ldots n$. If* Pr *satisfies CAR, then $S' \cdot \vec{\gamma}^T = \vec{1}^T$.*

(b) Let S' be a matrix consisting of a subset of the rows of S, and let $\mathcal{P}_{W,S'}$ be the set of distributions over W with support corresponding to S'; i.e.,

$$\mathcal{P}_{W,S'} = \{P_W \mid P_W(A) > 0 \textit{ iff } A \textit{ corresponds to a row in } S'\}.$$

If there exists a vector $\vec{\gamma} \geq \vec{0}$ such that $S' \cdot \vec{\gamma}^T = \vec{1}^T$, then, for all *$P_W \in \mathcal{P}_{W,S'}$, there exists a distribution* Pr *over $\mathcal{R}$ with $\Pr_W = P_W$ (i.e., the marginal of* Pr *on W is P_W) such that (a)* Pr *satisfies CAR and (b) $\Pr(X_O = U_j \mid X_W \in U_j) = \gamma_j$ for all j with $\Pr(X_W \in U_j) > 0$.*

Note that (b) is essentially a converse of (a). A natural question to ask is whether (b) would still hold if we replaced "for all $P_W \in \mathcal{P}_{W,S'}$ there exists Pr satisfying CAR with $\Pr_W = P_W$" by "for all distributions P_O over O there exists Pr satisfying CAR with $\Pr_O = P_O$." The answer is no; see Example 4.6(b)(ii).

Lemma 4.3 says that a distribution Pr that satisfies CAR and at the same time has $\Pr(X_W \in A) > 0$ for m different atoms A can exist if and only if a certain set of m linear equations in n unknowns has a solution. In many situations of interest, $m \geq n$ (note that m may be as large as $2^n - 1$). Not surprisingly then, in such situations there often can be *no* distribution Pr that satisfies CAR, as we show in the next subsection. On the other hand, if the set of equations $S'\vec{\gamma}^T = \vec{1}$ does have a solution in $\vec{\gamma}$, then the set of all solutions forms the intersection of an affine subspace (i.e. a hyperplane) of $\mathbf{R}^n$ and the positive orthant $[0, \infty)^n$. These solutions are just the conditional probabilities $\Pr(X_O = U_j \mid X_W \in U_j)$ for all distributions for which CAR holds that have support corresponding to S'. These conditional probabilities may then be extended to a distribution over $\mathcal{R}$ by setting $\Pr_W = P_W$ for an arbitrary distribution P_W over the worlds in atoms corresponding to S'; all Pr constructed in this way satisfy CAR.

Summarizing, we have the remarkable fact that for any given set of atoms $\mathcal{A}$ there are only two possibilities: either *no* distribution exists which has $\Pr(X_W \in A) > 0$ for all $A \in \mathcal{A}$ and satisfies CAR, or *for all* distributions P_W over worlds corresponding to atoms in A, there exists a distribution satisfying CAR with marginal distribution over worlds equal to P_W.

4.3 When CAR is Guaranteed *Not* to Hold

We now present a theorem that gives two explicit and easy-to-check sufficient conditions under which CAR cannot hold unless the probabilities of some atoms and/or observations are 0. The theorem is proved by showing that the condition of Lemma 4.3(a) cannot hold under the stated conditions.

We briefly recall some standard definitions from linear algebra. A set of vectors $\vec{v}_1, \ldots, \vec{v}_m$ is called *linearly dependent* if there exist coefficients $\lambda_1, \ldots, \lambda_m$ (not all zero) such that $\sum_{i=1}^m \lambda_i \vec{v}_i = \vec{0}$; the vectors are *affinely dependent* if there exist coefficients $\lambda_1, \ldots, \lambda_m$ (not all zero) such that $\sum_{i=1}^m \lambda_i \vec{v}_i = \vec{0}$ and $\sum_{i=1}^m \lambda_i = 0$. A vector $\vec{u}$ is called an *affine combination* of $\vec{v}_1, \ldots, \vec{v}_m$ if there exist coefficients $\lambda_1, \ldots, \lambda_m$ such that $\sum_{i=1}^m \lambda_i \vec{v}_i = \vec{u}$ and $\sum_{i=1}^m \lambda_i = 0$.

Theorem 4.4: *Let $\mathcal{R}$ be a set of runs over observations $O = \{U_1, \ldots, U_n\}$ and worlds W, and let S be the CARacterizing matrix for O and W.*

(a) Suppose that there exists a subset R of the rows in S and a vector $\vec{u} = (u_1, \ldots, u_n)$ that is an affine combination of the rows of R such that $u_j \geq 0$ for all $j \in \{1, \ldots, n\}$ and $u_{j^*} > 0$ for some $j^* \in \{1, \ldots, n\}$. Then there is no distribution $\Pr$ *on $\mathcal{R}$ that satisfies CAR such that* $\Pr(X_O = U_{j^*}) > 0$ *and* $\Pr(X_W \in A) > 0$ *for each $\mathcal{R}$-atom A corresponding to a row in R.*

(b) If there exists a subset R of the rows of S that is linearly dependent but not affinely dependent, then there is no distribution $\Pr$ *on $\mathcal{R}$ that satisfies CAR such that* $\Pr(X_W \in A) > 0$ *for each $\mathcal{R}$-atom A corresponding to a row in R.*

(c) Given a set R consisting of n linearly independent rows of S and a distribution P_W on W such that $P_W(A) > 0$ for all A corresponding to a row in R, there is a unique distribution P_O on O such that if $\Pr$ *is a distribution on $\mathcal{R}$ satisfying CAR and* $\Pr(X_W \in A) = P_W(A)$ *for each atom A corresponding to a row in R, then* $\Pr(X_O = U) = P_O(U)$.

It is well known that in an $m \times n$ matrix, at most n rows can be linearly independent. In many cases of interest (cf. Example 4.5 below), the number of atoms m is larger than the number of observations n, so that there must exist subsets R of rows of S that are linearly dependent. Thus, part (b) of Theorem 4.4 puts nontrivial constraints on the distributions that satisfy CAR.

The requirement in part (a) may seem somewhat obscure but it can be easily checked and applied in a number of situations, as illustrated in Example 4.5 and 4.6 below. Part (c) says that in many other cases of interest where neither part (a) nor (b) applies, even if a distribution on $\mathcal{R}$ exists satisfying CAR, the probabilities of making the observations are completely determined by the probability of various events in the world occurring, which seems rather unreasonable.

Example 4.5: Consider the CARacterizing matrix of Example 4.2. Notice there exists an affine combination of the first two rows that is not $\vec{0}$ and has no negative components:

$$-1 \cdot \begin{pmatrix} 1 \\ 0 \end{pmatrix} + 1 \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

Similarly, there exists an affine combination of the last two rows that is not $\vec{0}$ and has no negative components. It follows from Theorem 4.4(a) that there is no distribution satisfying CAR that gives both of the observations $X_O = U_1$ and $X_O = U_2$ positive probability and either (a) gives both $X_W \in U_1 - U_2$ and $X_W \in U_1 \cap U_2$ positive probability or (b) gives both $X_W \in U_2 - U_1$ and $X_W \in U_1 \cap U_2$ positive probability. If both observations have positive probability, then CAR can hold only if the probability of $U_1 \cap U_2$ is either 0 or 1. (Example 3.3 already shows this using a more direct argument.) ∎

The next example further illustrates that in general, it can be very difficult to satisfy CAR.

Example 4.6: Suppose that $O = \{U_1, U_2, U_3\}$, and all three observations can be made with positive probability. It turns out that in this situation the CAR condition can hold, but only if (a) $\Pr(X_W \in U_1 \cap U_2 \cap U_3) = 1$ (i.e., all of U_1, U_2, and U_3 must hold), (b) $\Pr(X_W \in ((U_1 \cap U_2) - U_3) \cup ((U_2 \cap U_3) - U_1) \cup ((U_1 \cap U_3) - U_2)) = 1$ (i.e., exactly two of U_1, U_2, and U_3 must hold), (c) $\Pr(X_W \in (U_1 - (U_2 \cup U_3)) \cup (U_2 - (U_1 \cup U_3)) \cup (U_3 - (U_2 \cup U_1))) = 1$ (i.e., exactly one of U_1, U_2, or U_3 must hold), or (d) one of $(U_1 - (U_2 \cup U_3)) \cup (U_2 \cap U_3)$, $(U_2 - (U_1 \cup U_3)) \cup (U_1 \cap U_3)$ or $(U_3 - (U_1 \cup U_2)) \cup (U_1 \cap U_2)$ has probability 1 (either exactly one of U_1, U_2, or U_3 holds, or the remaining two both hold).

We first check that CAR can hold in all these cases. It should be clear that CAR can hold in case (a). Moreover, there are no constraints on $\Pr(X_O = U_i \mid X_W = w)$ for $w \in U_1 \cap U_2 \cap U_3$ (except, by the CAR condition, for each fixed i, the probability must be the same for all $w \in U_1 \cap U_2 \cap U_3$, and the three probabilities must sum to 1).

For case (b), let A_i be the atom where exactly two of U_1, U_2, and U_3 hold, and U_i does not hold, for $i = 1, 2, 3$. Suppose that $\Pr(X_W \in A_1 \cup A_2 \cup A_3) = 1$. Note that, since all three observations can be made with positive probability, at least two of A_1, A_2, and A_3 must have positive probability. Hence we can distinguish between two subcases: (i) only two of them have positive probability, and (ii) all three have positive probability.

For subcase (i), suppose without loss of generality that only A_1 and A_2 have positive probability. Then it immediately follows from the CAR condition that there must be some α with $0 < \alpha < 1$ such that $\Pr(X_O = U_3 \mid X_W = w) = \alpha$, for all $w \in A_1 \cup A_2$ such that $\Pr(X_W = w) > 0$. Thus, $\Pr(X_O = U_1 \mid X_W = w) = 1 - \alpha$ for all $w \in A_2$ such that $\Pr(X_W = w) > 0$, and $\Pr(X_O = U_2 \mid X_W = w) = 1 - \alpha$ for all $w \in A_1$ such that $\Pr(X_W = w) > 0$.

Subcase (ii) is more interesting. The rows of the CARacterizing matrix S corresponding to A_1, A_2, and A_3 are (0 1 1), (1 0 1), and (1 1 0), respectively. Now Lemma 4.3(a) tells us that if Pr satisfies CAR, then we must have $S \cdot \vec{\gamma}^T = \vec{1}^T$ for some $\vec{\gamma} = (\gamma_1, \gamma_2, \gamma_3)$ with $\gamma_i = \Pr(X_O = U_i \mid X_W \in U_i)$. These three linear equations have solution

$$\gamma_1 = \gamma_2 = \gamma_3 = \frac{1}{2}.$$

Since this solution is unique, it follows by Lemma 4.3(b) that *all* distributions that satisfy CAR must have conditional probabilities $\Pr(X_O = U_i \mid X_W \in U_i) = 1/2$, and that their marginal distributions on W can be arbitrary. This fully characterizes the set of distributions Pr for which CAR holds in this case. Note that for $i = 1, 2, 3$, since we can write $\gamma_i = \Pr(X_O = U_i)/\Pr(X_W \in U_i)$ we have $\Pr(X_O = U_i) = \Pr(X_w \in U_i)\gamma_i \leq 1/2$ so that, in contrast to the marginal distribution over W, the marginal distribution over O cannot be chosen arbitrarily.

In case (c), it should also be clear that CAR can hold. Moreover, $\Pr(X_0 = U_i \mid X_W = w)$ is either 0 or 1, depending on whether $w \in U_i$. Finally, for case (d), suppose that $\Pr(X_W \in U_1 \cup (U_2 \cap U_3)) = 1$. CAR holds iff there exists α such that $\Pr(X_O = U_2 \mid X_W = w) = \alpha$ and $\Pr(X_O = U_3 \mid X_W = w) = 1 - \alpha$ for all $w \in U_2 \cap U_3$ such that $\Pr(X_W = w) > 0$. (Of course, $\Pr(X_O = U_1 \mid X_W = w) = 1$ for all $w \in U_1$ such that $\Pr(X_W = w) > 0$.)

Now we show that CAR cannot hold in any other case. First suppose that $0 < \Pr(X_W \in U_1 \cap U_2 \cap U_3) < 1$. Thus, there must be at least one other atom A such that $\Pr(X_W \in A) > 0$. The row corresponding to the atom $U_1 \cap U_2 \cap U_3$ is $(1\ 1\ 1)$. Suppose r is the row corresponding to the other atom A. Since S is a 0-1 matrix, the vector $(1\ 1\ 1) - r$ gives is an affine combination of $(1\ 1\ 1)$ and r that is nonzero and has nonnegative components. It now follows by Theorem 4.4 that CAR cannot hold in this case.

Similar arguments give a contradiction in all the other cases; we leave details to the reader. ∎

4.4 Discussion: "CAR is Everything" vs. "Sometimes CAR is Nothing"

In one of their main theorems, Gill, van der Laan, and Robins (1997, Section 2) show that the CAR assumption is untestable from observations of X_O alone, in the sense that the assumption "Pr satisfies CAR" imposes no restrictions at all on the marginal distribution $\Pr_O$ on X_O. More precisely, they show that for every finite set W of worlds, every set O of observations, and every distribution P_O on O, there is a distribution $\Pr^*$ on $\mathcal{R}$ such that $\Pr^*_O$ (the marginal of $\Pr^*$ on O) is equal to P_O and $\Pr^*$ satisfies CAR. The authors summarize this as "CAR is everything".

We must be careful in interpreting this result. Theorem 4.4 shows that, for many combinations of O and W, CAR can hold *only* for distributions Pr with $\Pr(X_W \in A) = 0$ for some atoms A. (In the previous sections, we called such distributions "degenerate".) In our view, this says that in some cases, CAR effectively cannot hold. To see why, first suppose we are given a set W of worlds and a set O of observations. Now we may feel confident *a priori* that some $U_0 \in O$ and some $w_0 \in W$ cannot occur in practice. In this case, we are willing to consider only distributions Pr on $O \times W$ that have $\Pr(X_O = U_0) = 0$, $\Pr(X_W = w_0) = 0$. (For example, W may be a product space $W = W_a \times W_b$ and it is known that some combination $w_a \in W_a$ and w_b in W_b can never occur together; then $\Pr(X_w = (w_a, w_b)) = 0$.) Define O^* to be the subset of O consisting of all U that we cannot a priori rule out; similarly, W^* is the subset of W consisting of all w that we cannot a priori rule out. By Theorem 4.4, it is still possible that O^* and W^* are such that, even if we restrict to runs where only observations in O^* are made, CAR can only hold if $\Pr(X_W \in A) = 0$ for some atoms (nonempty subsets) $A \subseteq W^*$. This means that CAR may force us to assign probability 0 to some events that, *a priori*, were considered possible. Examples 3.3 and 4.6 illustrate this phenomenon. We may summarize this as "sometimes CAR is nothing".

Given therefore that CAR imposes such strong conditions, the reader may wonder why there is so much study of the CAR condition in the statistics literature. The reason is that some of the special situations in which CAR holds often arise in missing data and survival analysis problems. Here is an example: Suppose that the set of observations can be written as $O = \cup_{i=1}^{k} \Pi_i$, where each Π_i is a partition of W (that is, a set of pairwise disjoint subsets of W whose union is W). Further suppose that observations are generated by the following process, which we call CARgen. Some i between 1 and k is chosen according to some arbitrary distribution P_0; independently, $w \in W$ is chosen according to P_W. The agent then observes the unique $U \in \Pi_i$ such that $w \in U$. Intuitively, the partitions Π_i may represent the observations that can be made with a particular sensor. Thus, P_0 determines the probability that a particular sensor is chosen; P_W determines the probability that a particular world is chosen. The sensor and the world together determine the observation that is made. It is easy to see that this mechanism induces a distribution on $\mathcal{R}$ for which CAR holds.

The special case with $O = \Pi_1 \cup \Pi_2$, $\Pi_1 = \{W\}$, and $\Pi_2 = \{\{w\} \mid w \in W\}$ corresponds to a simple missing data problem (Example 4.7 below). Intuitively, either complete information is given, or there is no data at all. In this context, CAR is often called *MAR: missing at random*. In more realistic MAR problems, we may observe a vector with some of its components missing. In such cases the CAR condition sometimes still holds. In practical missing data problems, the goal is often to infer the distribution Pr on runs $\mathcal{R}$ from successive observations of X_O. That is, one observes a sample $U_{(1)}, U_{(2)}, \ldots, U_{(n)}$, where $U_{(i)} \in O$. Typically, the $U_{(i)}$ are assumed to be an i.i.d. (independently identically distributed) sample of outcomes of X_O. The corresponding "worlds" $w_1, w_2, \ldots$ (outcomes of X_W) are *not* observed. Depending on the situation, Pr may be completely unknown or is assumed to be a member of some parametric family of distributions. If the number of observations n is large, then clearly the sample $U_{(1)}, U_{(2)}, \ldots, U_{(n)}$ can be used to obtain a reasonable estimate of $\Pr_O$, the marginal distribution on X_O. But one is interested in the full distribution Pr. That distribution usually cannot be inferred without making additional assumptions, such as the CAR assumption.

Example 4.7: (adapted from (Scharfstein, Daniels, & Robins, 2002)) Suppose that a medical study is conducted to test the effect of a new drug. The drug is administered to a group of patients on a weekly basis. Before the experiment is started and after it is finished, some characteristic (say, the blood pressure) of the patients is measured. The data are thus differences in blood pressure for individual patients before and after the treatment. In practical studies of this kind, often several of the patients drop out of the experiment. For such patients there is then no data. We model this as follows: W is the set of possible values of the characteristic we are interested in (e.g., blood pressure difference). $O = \Pi_1 \cup \Pi_2$ with $\Pi_1 = \{W\}$, and $\Pi_2 = \{\{w\} \mid w \in W\}$ as above. For "compliers" (patients that did not drop out), we observe $X_O = \{w\}$, where w is the value of the characteristic we want to measure. For dropouts, we observe $X_O = W$ (that is, we observe nothing at all). We thus have, for example, a sequence of observations $U_1 = \{w_1\}, U_2 = \{w_2\}, U_3 = W, U_4 = \{w_4\}, U_5 = W, \ldots, U_n = \{w_n\}$. If this sample is large enough, we can use it to obtain a reasonable estimate of the probability that a patient drops out (the ratio of outcomes with

$U_i = W$ to the total number of outcomes). We can also get a reasonable estimate of the distribution of X_W for the complying patients. Together these two distributions determine the distribution of X_O.

We are interested in the effect of the drug in the general population. Unfortunately, it may be the case that the effect on dropouts is different from the effect on compliers. (Scharfstein, Daniels, and Robins (2002) discuss an actual medical study in which physicians judged the effect on dropouts to be very different from the effect of compliers.) Then we cannot infer the distribution on W from the observations $U_1, U_2, \ldots$ alone without making additional assumptions about how the distribution for dropouts is related to the distribution for compliers. Perhaps the simplest such assumption that one can make is that the distribution of X_W for dropouts *is* in fact the same as the distribution of X_W for compliers: the data are "missing at random". Of course, this assumption is just the CAR assumption. By Theorem 3.1(a), CAR holds iff for all $w \in W$

$$\Pr(X_W = w \mid X_O = W) = \Pr(X_W = w \mid X_W \in W) = \Pr(X_W = w),$$

which means just that the distribution of W is independent of whether a patient drops out ($X_O = W$) or not. Thus, *if* CAR can be assumed, then we can infer the distribution on W (which is what we are really interested in). ∎

Many problems in missing data and survival analysis are of the kind illustrated above: The analysis would be greatly simplified if CAR holds, but whether or not this is so is not clear. It is therefore of obvious interest to investigate whether, from observing the "coarsened" data $U_{(1)}, U_{(2)}, \ldots, U_{(n)}$ alone, it may already be possible to test the assumption that CAR holds. For example, one might imagine that there are distributions on X_O for which CAR simply cannot hold. If the empirical distribution of the U_i were "close" (in the appropriate sense) to a distribution that rules out CAR, the statistician might infer that Pr does not satisfy CAR. Unfortunately, if O is finite, then the result of Gill, van der Laan, and Robins (1997, Section 2) referred to at the beginning of this section shows that we can never rule out CAR in this way.

We are interested in the question of whether CAR can hold in a "nondegenerate" sense, given O and W. From this point of view, the slogan "sometimes CAR is nothing" makes sense. In contrast, Gill et al. (1997) were interested in the question whether CAR can be tested from observations of X_O alone. From that point of view, the slogan "CAR is everything" makes perfect sense. In fact, Gill, van der Laan, and Robins were quite aware, and explicitly stated, that CAR imposes very strong assumptions on the distribution Pr. In a later paper, it was even implicitly stated that in some cases CAR forces $\Pr(X_W \in A) = 0$ for some atoms A (Robins, Rotnitzky, & Scharfstein, 1999, Section 9.1). Our contribution is to provide the precise conditions (Lemma 4.3 and Theorem 4.4) under which this happens.

Robins, Rodnitzky, and Scharfstein (1999) also introduced a Bayesian method (later extended by Scharfstein et al. (2002)) that allows one to specify a prior distribution over a parameter α which indicates, in a precise sense, how much Pr deviates from CAR. For example, $\alpha = 0$ corresponds to the set of distributions Pr satisfying CAR. The precise connection between this work and ours needs further investigation.

4.5 A Mechanism for Generating Distributions Satisfying CAR

In Theorem 3.1 and Lemma 4.3 we described CAR in an algebraic way, as a collection of probabilities satisfying certain equalities. Is there a more "procedural" way of representing CAR? In particular, does there exist a single mechanism that gives rise to CAR such that *every* case of CAR can be viewed as a special case of this mechanism? We have already encountered a possible candidate for such a mechanism: the CARgen procedure. In Section 4.4 we described this mechanism and indicated that it generates only distributions that satisfy CAR. Unfortunately, as we now show, there exist CAR distributions that cannot be interpreted as being generated by CARgen.

Our example is based on an example given by Gill, van der Laan, and Robins (1997), who were actually the first to consider whether there exist 'natural' mechanisms that generate all and only distributions satisfying CAR. They show that in several problems of survival analysis, observations are generated according to what they call a *randomized monotone coarsening scheme*. They also show that their randomized scheme generates only distributions that satisfy CAR. In fact, the randomized monotone coarsening scheme turns out to be a special case of CARgen, although we do not prove this here. Gill, van der Laan, and Robins show by example that the randomized coarsening schemes do not suffice to generate all CAR distributions. Essentially the same example shows that CARgen does not either.

Example 4.8: Consider subcase (ii) of Example 4.6 again. Let U_1, U_2, U_3 and A_1, A_2 and A_3 be as in that example, and assume for simplicity that $W = A_1 \cup A_2 \cup A_3$. The example showed that there exists distributions Pr satisfying CAR in this case with $\Pr(A_i) > 0$ for $i \in \{1, 2, 3\}$, all having conditional probabilities $\Pr(X_O = U_i \mid X_W = w) = 1/2$ for all $w \in U_i$. Clearly, U_1, U_2 and U_3 cannot be grouped together to form a set of partitions of W. So, even though CAR holds for Pr, CARgen cannot be used to simulate Pr. ∎

While Gill, van der Laan, and Robins (1997, Section 3) ask whether there exists a general mechanism for generating all and only CAR distributions, they do not make this question mathematically precise. As noted by Gill, van der Laan, and Robins, the problem here is that without any constraint to what constitutes a valid mechanism, there is clearly a trivial solution to the problem: Given a distribution Pr satisfying CAR, we simply draw a world w according to $\Pr_W$, and then draw U such that $w \in U$ according to the distribution $\Pr(X_O = U \mid X_W = w)$. This is obviously cheating in some sense. Intuitively, it seems that a 'reasonable' mechanism should not be allowed to choose U according to a distribution depending on w. It does not have that kind of control over the observations that are made.

So what counts as a "reasonable" mechanism? Intuitively, the mechanism should be able to control only what can be controlled in an experimental setup. We can think of the mechanism as an "agent" that uses a set of sensors to obtain information about the world. The agent does not have control over P_W. While the agent may certainly choose which sensor to use, it is not reasonable to assume that she can control their output (or exactly what they can sense). Indeed, given a world w, the observation returned by the sensor is fully determined. This is exactly the setup implemented in the CARgen scheme discussed in Section 4.4, which we therefore regard as a legitimate mechanism. We now introduce a procedure CARgen*, which extends CARgen, and turns out to generate all and only distributions satisfying CAR. Just like CARgen, CARgen* assumes that there is a collection of sensors, and it consults a given sensor with a certain predetermined probability.

However, unlike CARgen, CARgen* may ignore a sensor reading. The decision whether or not a sensor reading is ignored is allowed to depend only on the sensor that has been chosen by the agent, and on the observation generated by that sensor. It is not allowed to depend on the actual world, since the agent may not know what the actual world is. Once again, the procedure is "reasonable" in that the agent is only allowed to control what can be controlled in an experimental setup.

We now present CARgen*, then argue that it is "reasonable" in the sense above, and that it generates all and only CAR distributions.

Procedure CARgen*

1. Preparation:

   - Fix an arbitrary distribution P_W on W.
   - Fix a set $\mathcal{P}$ of partitions of W, and fix an arbitrary distribution $P_{\mathcal{P}}$ on $\mathcal{P}$.
   - Choose numbers $q \in [0,1)$ and $q_{U|\Pi} \in [0,1]$ for each pair (U,Π) such that $\Pi \in \mathcal{P}$ and $U \in \Pi$ satisfying the following constraint, for each $w \in W$ such that $P_W(w) > 0$:
   $$q = \sum_{\{(U,\Pi):\, w \in U,\, U \in \Pi\}} P_{\mathcal{P}}(\Pi) q_{U|\Pi}. \tag{4}$$

2. Generation:

   2.1 Choose $w \in W$ according to P_W.

   2.2 Choose $\Pi \in \mathcal{P}$ according to $P_{\mathcal{P}}$. Let U be the unique set in Π such that $w \in U$.

   2.3 With probability $1 - q_{U|\Pi}$, return (w, U) and halt. With probability $q_{U|\Pi}$, go to step 2.2.

It is easy to see that CARgen is the special case of CARgen* where $q_{U|\Pi} = 0$ for all (U,Π). Allowing $q_{U|\Pi} > 0$ gives us a little more flexibility. To understand the role of the constraint (4), note that $q_{U|\Pi}$ is the probability that the algorithm does not terminate at step 2.3, given that U and Π are chosen at step 2.2. It follows that the probability q_w that a pair (w, U) is not output at step 2.3 for some U is

$$q_w = \sum_{\{(U,\Pi):\, w \in U,\, U \in \Pi\}} P_{\mathcal{P}}(\Pi) q_{U|\Pi}.$$

Thus, (4) says that the probability q_w that a pair whose first component is w is not output at step 2.3 is the same for all $w \in W$.

CARgen* can generate the CAR distribution in Example 4.8, which could not be generated by CARgen. To see this, using the same notation as in the example, consider the set of partitions $\mathcal{P} = \{\Pi_1, \Pi_2, \Pi_3\}$ with $\Pi_i = \{U_i, A_i\}$. Let $P_{\mathcal{P}}(\Pi_1) = P_{\mathcal{P}}(\Pi_2) = P_{\mathcal{P}}(\Pi_3) = 1/3$, $q_{U_i|\Pi_i} = 0$, and $q_{A_i|\Pi_i} = 1$. It is easy to verify that for all $w \in W$, we have that $\sum_{\{U,\Pi: w \in U, U \in \Pi\}} P_{\mathcal{P}}(\Pi) q_{U|\Pi} = 1/3$, so that the constraint (4) is satisfied. Moreover, direct calculation shows that, for arbitrary P_W, the distribution $\Pr^*$ on runs generated by CARgen* with this choice of parameters is precisely the unique distribution satisfying CAR in this case.

So why is CARgen* reasonable? Even though we have not given a formal definition of "reasonable" (although it can be done in the runs framework—essentially, each step of the algorithm can depend only on information available to the experimenter, where the "information" is encoded in the observations made by the experimenter in the course of running the algorithm), it is not hard to see that CARgen* satisfies our intuitive desiderata. The key point is that all the relevant steps in the algorithm can be carried out by an experimenter. The parameters q and $q_{U|\Pi}$ for $\Pi \in \mathcal{P}$ and $U \in \Pi$ are chosen before the algorithm begins; this can certainly be done by an experimenter. Similarly, it is straightforward to check that the equation (4) holds for each $w \in W$. As for the algorithm itself, the experimenter has no control over the choice of w; this is chosen by nature according its distribution, P_W. However, the experimenter can perform steps 2.2 and 2.3, that is choosing $\Pi \in \mathcal{P}$ according to the probability distribution $P_{\mathcal{P}}$, and rejecting the observation U with probability $q_{U|\Pi}$, since the experimenter knows both the sensor chosen (i.e., Π) and the observation (U).

The following theorem shows that CARgen* does exactly what we want.

Theorem 4.9: *Given a set $\mathcal{R}$ of runs over a set W of worlds and a set O of observations,* Pr *is a distribution on $\mathcal{R}$ that satisfies CAR if and only if there is a setting of the parameters in (step 1 of)* CARgen* *such that, for all $w \in W$ and $U \in O$,* $\Pr(\{r : X_W(r) = w,\ X_O(r) = U\})$ *is the probability that* CARgen* *returns* (w, U).

5. Beyond Observations of Events

In the previous section, we assumed that the information received was of the form "the actual world is in U". Information must be in this form to apply conditioning. But in general, information does not always come in such nice packages. In this section we study more general types of observations, leading to generalizations of conditioning.

5.1 Jeffrey Conditioning

Perhaps the simplest possible generalization is to assume that there is a partition $\{U_1, \ldots, U_n\}$ of W and the agent observes $\alpha_1 U_1; \ldots; \alpha_n U_n$, where $\alpha_1 + \cdots + \cdots \alpha_n = 1$. This is to be interpreted as an observation that leads the agent to believe U_j with probability α_j, for $j = 1, \ldots, n$. According to Jeffrey conditioning, given a distribution P_W on W,

$$P_W(V \mid \alpha_1 U_1; \ldots; \alpha_n U_n) = \alpha_1 P_W(V \mid U_1) + \cdots + \alpha_n P_W(V \mid U_n).$$

Jeffrey conditioning is defined only if $\alpha_i > 0$ implies that $P_W(U_i) > 0$; if $\alpha_i = 0$ and $P_W(U_i) = 0$, then $\alpha_i P_W(V \mid U_i)$ is taken to be 0. Clearly ordinary conditioning is the special case of Jeffrey conditioning where $\alpha_i = 1$ for some i so, as is standard, we deliberately use the same notation for updating using Jeffrey conditioning and ordinary conditioning.

We now want to determine when updating in the naive space using Jeffrey conditioning is appropriate. Thus, we assume that the agent's observations now have the form of $\alpha_1 U_1; \ldots; \alpha_n U_n$ for some partition $\{U_1, \ldots, U_n\}$ of W. (Different observations may, in general, use different partitions.) Just as we did for the case that observations are events (Section 3, first paragraph), we once again assume that the agent's observations are accurate. What does that mean in the present context? We simply require that, conditional on making the observation, the probability of U_i really is α_i for $i = 1, \ldots, n$. That is, for $i = 1, \ldots, n$, we have

$$\Pr(X_W \in U_i \mid X_O = \alpha_1 U_1; \ldots; \alpha_n U_n) = \alpha_i. \tag{5}$$

This clearly generalizes the requirement of accuracy given in the case that the observations are events.

Not surprisingly, there is a generalization of the CAR condition that is needed to guarantee that Jeffrey conditioning can be applied to the naive space.

Theorem 5.1: *Fix a probability* $\Pr$ *on* $\mathcal{R}$*, a partition* $\{U_1, \ldots, U_n\}$ *of* W*, and probabilities* $\alpha_1, \ldots, \alpha_n$ *such that* $\alpha_1 + \cdots + \alpha_n = 1$*. Let* C *be the observation* $\alpha_1 U_1; \ldots; \alpha_n U_n$*. Fix some* $i \in \{1, \ldots, n\}$*. Then the following are equivalent:*

(a) If $\Pr(X_O = C) > 0$*, then* $\Pr(X_W = w \mid X_O = C) = \Pr_W(w \mid \alpha_1 U_1; \ldots; \alpha_n U_n)$ *for all* $w \in U_i$*.*

(b) $\Pr(X_O = C \mid X_W = w) = \Pr(X_O = C \mid X_W \in U_i)$ *for all* $w \in U_i$ *such that* $\Pr(X_W = w) > 0$*.*

Part (b) of Theorem 5.1 is analogous to part (c) of Theorem 3.1. There are a number of conditions equivalent to (b) that we could have stated, similar in spirit to the conditions in Theorem 3.1. Note that these are even more stringent conditions than are required for ordinary conditioning to be appropriate.

Examples 3.3 and 4.6 already suggest that there are not too many nontrivial scenarios where applying Jeffrey conditioning to the naive space is appropriate. However, just as for the original CAR condition, there do exist special situations in which generalized CAR is a realistic assumption. For ordinary CAR, we mentioned the CARgen mechanism (Section 4.5). For Jeffrey conditioning, a similar mechanism may be a realistic model in some situations where all observations refer to the same partition $\{U_1, \ldots, U_n\}$ of W. We now describe a scenario for such a situation. Suppose O consists of $k > 1$ observations $C_1, \ldots, C_k$ with $C_i = \alpha_{i1} U_1; \ldots; \alpha_{in} U_n$ such that all $\alpha_{ij} > 0$. Now, fix n (arbitrary) conditional distributions $\Pr_j$, $j = 1, \ldots, n$, on W. Intuitively, $\Pr_j$ is $\Pr_W(\cdot \mid U_j)$. Consider the following mechanism: first an observation C_i is chosen (according to some distribution P_O on O); then a set U_j is chosen with probability α_{ij} (i.e., according to the distribution induced by C_i); finally, a world $w \in U_j$ is chosen according to $\Pr_j$.

If the observation C_i and world w are generated this way, then the generalized CAR condition holds, that is, conditioning in the sophisticated space coincides with Jeffrey conditioning:

Proposition 5.2: *Consider a partition* $\{U_1, \ldots, U_n\}$ *of* W *and a set of* $k > 1$ *observations* O *as above. For every distribution* P_O *on* O *with* $P_O(C_i) > 0$ *for all* $i \in \{1, \ldots, k\}$*, there exists a distribution* $\Pr$ *on* $\mathcal{R}$ *such that* $P_O = \Pr_O$ *(i.e.* P_O *is the marginal of* $\Pr$ *on* O*) and* $\Pr$ *satisfies the generalized CAR condition (Theorem 5.1(b)) for* $U_1, \ldots, U_n$*.*

Proposition 5.2 demonstrates that, even though the analogue of the CAR condition expressed in Theorem 5.1 is hard to satisfy in general, at least if the set $\{U_1, \ldots, U_n\}$ is the same for all observations, then for every such set of observations there exist *some* priors Pr on $\mathcal{R}$ for which the CAR-analogue *is* satisfied for all observations. As we show next, for MRE updating, this is no longer the case.

5.2 Minimum Relative Entropy Updating

What about cases where the constraints are not in the special form where Jeffrey's conditioning can be applied? Perhaps the most common approach in this case is to use MRE. Given a constraint (where a constraint is simply a set of probability distributions—intuitively, the distributions satisfying the constraint) and a prior distribution P_W on W, the idea is to pick, among all distributions satisfying the constraint, the one that is "closest" to the prior distribution, where the "closeness" of P'_W to P_W is measured using relative entropy. The *relative entropy between* P'_W *and* P_W (Kullback & Leibler, 1951; Cover & Thomas, 1991) is defined as

$$\sum_{w \in W} P'_W(w) \log \left(\frac{P'_W(w)}{P_W(w)} \right).$$

(The logarithm here is taken to the base 2; if $P'_W(w) = 0$ then $P'_W(w) \log(P'_W(w)/P_W(w))$ is taken to be 0. This is reasonable since $\lim_{x \to 0} x \log(x/c) = 0$ if $c > 0$.) The relative entropy is finite provided that P'_W is *absolutely continuous* with respect to P_W, in that if $P_W(w) = 0$, then $P'_W(w) = 0$, for all $w \in W$. Otherwise, it is defined to be infinite.

The constraints we consider here are all closed and convex sets of probability measures. In this case, it is known that there is a unique distribution that satisfies the constraints and minimizes the relative entropy. Given a nonempty constraint C and a probability distribution P_W on W, let $P_W(\cdot \mid C)$ denote the distribution that minimizes relative entropy with respect to P_W.

If the constraints have the form to which Jeffrey's Rule is applicable, that is, if they have the form $\{P'_W : P'_W(U_i) = \alpha_i, i = 1, \ldots, n\}$ for some partition $\{U_1, \ldots, U_n\}$, then it is well known that the distribution that minimizes entropy relative to a prior P_W is $P_W(\cdot \mid \alpha_1 U_1; \ldots; \alpha_n U_n)$ (see, e.g., (Diaconis & Zabell, 1986)). Thus, MRE updating generalizes Jeffrey conditioning (and hence also standard conditioning).

To study MRE updating in our framework, we assume that the observations are now arbitrary closed convex constraints on the probability measure. Again, we assume that the observations are accurate in that, conditional on making the observation, the constraints hold. For now, we focus on the simplest possible case that cannot be handled by Jeffrey updating. In this case, constraints (observations) still have the form $\alpha_1 U_1; \ldots; \alpha_n U_n$, but now the U_i's do not have to form a partition (they may overlap and/or not cover W) and the α_i do not have to sum to 1. Such an observation is accurate if it satisfies (5), just as before.

We can now ask the same questions that we asked before about ordinary conditioning and Jeffrey conditioning in the naive space.

1. Is there an alternative characterization of the conditions under which MRE updating coincides with conditioning in the sophisticated space? That is, are there analogues of Theorem 3.1 and Theorem 5.1 for MRE updating?

2. Are there combinations of O and W for which it is not even possible that MRE can coincide with conditioning in the sophisticated space?

With regard to question 1, it is easy to provide a counterexample showing that there is no obvious analogue to Theorem 5.1 for MRE. There is a constraint C such that the condition of part (a) of Theorem 5.1 holds for MRE updating whereas part (b) does not hold. (We omit the details here.) Of course, it is possible that there are some quite different conditions that characterize when MRE updating coincides with conditioning in the sophisticated space. However, even if they exist, such conditions may be uninteresting in that they may hardly ever apply. Indeed, as a partial answer to question 2, we now introduce a very simple setting in which MRE updating necessarily leads to a result different from conditioning in the sophisticated space.

Let U_1 and U_2 be two subsets of W such that $V_1 = U_1 - U_2$, $V_2 = U_2 - U_1$, $V_3 = U_1 \cap U_2$, and $V_4 = W - (U_1 \cup U_2)$ are all nonempty. Consider a constraint of the form $C = \alpha_1 U_1; \alpha_2 U_2$, where α_1, α_2 are both in $(0,1)$. We investigate what happens if we use MRE updating on C. Since U_1 and U_2 overlap and do not cover the space, in general Jeffrey conditioning cannot be applied to update on C. There are some situations where, despite the overlap, Jeffrey conditioning can essentially be applied. We say that observation $C = \alpha_1 U_1; \alpha_2 U_2$ is *Jeffrey-like* iff, after MRE updating on one of the constraints $\alpha_1 U_1$ or $\alpha_2 U_2$, the other constraint holds as well. That is, C is Jeffrey-like (with respect to P_W) if either $P_W(U_2 \mid \alpha_1 U_1) = \alpha_2$ or $P_W(U_1 \mid \alpha_2 U_2) = \alpha_1$. Suppose that $P_W(U_2 \mid \alpha_1 U_1) = \alpha_2$; then it is easy to show that $P_W(\cdot \mid \alpha_1 U_1) = P_W(\cdot \mid \alpha_1 U_1; \alpha_2 U_2)$.

Intuitively, if the "closest" distribution P'_W to P_W that satisfies $P'_W(U_1) = \alpha_1$ also satisfies $P'_W(U_2) = \alpha_2$, then P'_W is the closest distribution to P_W that satisfies the constraint $C = \alpha_1 U_1; \alpha_2 U_2$. Note that MRE updating on αU is equivalent to Jeffrey conditioning on $\alpha U; (1-\alpha)(W - U)$. Thus, if C is Jeffrey-like, then updating with C is equivalent to Jeffrey updating.

Theorem 5.3: *Given a set $\mathcal{R}$ of runs and a set $O = \{C_1, C_2\}$ of observations, where $C_i = \alpha_{i1} U_1; \alpha_{i2} U_2$, for $i = 1, 2$, let $\Pr$ be a distribution on $\mathcal{R}$ such that $\Pr(X_O = C_1)$, $\Pr(X_O = C_2) > 0$, and $\Pr_W(w) = \Pr(X_W = w) > 0$ for all $w \in W$. Let $\Pr^i = \Pr(\cdot \mid X_O = C_i)$, and let $\Pr^i_W$ be the marginal of $\Pr^i$ on W. If either C_1 or C_2 is not Jeffrey-like, then we cannot have $\Pr^i_W = \Pr_W(\cdot \mid C_i)$, for both $i = 1, 2$.*

For fixed U_1 and U_2, we can identify an observation $\alpha_1 U_1; \alpha_2 U_2$ with the pair $(\alpha_1, \alpha_2) \in (0,1)^2$. Under our conditions on U_1 and U_2, the set of all Jeffrey-like observations is a subset of 0 (Lebesgue) measure of this set. Thus, the set of observations for which MRE conditioning corresponds to conditioning in the sophisticated space is a (Lebesgue) measure 0 set in the space of possible observations. Note however, that this set depends on the prior P_W over W.

A result similar to Theorem 5.3 was proved by Seidenfeld (1986) (and considerably generalized by Dawid (2001)). Seidenfeld shows that, under very weak conditions, MRE updating cannot coincide with sophisticated conditioning if the observations have the form "the conditional probability of U given V is α" (as is the case in the Judy Benjamin problem). Theorem 5.3 shows that this is impossible even for observations of the much simpler form $\alpha_1 U_1; \alpha_2 U_2$, unless we can reduce the problem to Jeffrey conditioning (in which case Theorem 5.1 applies).

observation type	*set of observations O*	*simplest applicable update rule*	*when it coincides with sophisticated conditioning*
event	pairwise disjoint	naive conditioning	always (Proposition 4.1)
event	arbitrary set of events	naive conditioning	iff CAR holds (Theorem 3.1)
probability vector	probabilities of partition	Jeffrey conditioning	iff generalization of CAR holds (Theorem 5.1)
probability vector	probabilities of two overlapping sets	MRE	if both observations Jeffrey-like (Theorem 5.3)

Figure 1: Conditions under which updating in the naive space coincides with conditioning in the sophisticated space.

6. Discussion

We have studied the circumstances under which ordinary conditioning, Jeffrey conditioning, and MRE updating in a naive space can be justified, where "justified" for us means "agrees with conditioning in the sophisticated space". The main message of this paper is that, except for quite special cases, the three methods cannot be justified. Figure 1 summarizes the main insights of this paper in more detail.

As we mentioned in the introduction, the idea of comparing an update rule in a "naive space" with conditioning in a "sophisticated space" is not new; it appears in the CAR literature and the MRE literature (as well as in papers such as (Halpern & Tuttle, 1993) and (Dawid & Dickey, 1977)). In addition to bringing these two strands of research together, our own contributions are the following: (a) we show that the CAR framework can be used as a general tool to clarify many of the well-known paradoxes of conditional probability; (b) we give a general characterization of CAR in terms of a binary-valued matrix, showing that in many realistic scenarios, the CAR condition *cannot* hold (Theorem 4.4); (c) we define a mechanism CARgen* that generates all and only distributions satisfying CAR (Theorem 4.9); (d) we show that the CAR condition has a natural extension to cases where Jeffrey conditioning can be applied (Theorem 5.1); and (e) we show that no CAR-like condition can hold in general for cases where only MRE (and not Jeffrey) updating can be applied (Theorem 5.3).

Our results suggest that working in the naive space is rather problematic. On the other hand, as we observed in the introduction, working in the sophisticated space (even assuming it can be constructed) is problematic too. So what are the alternatives?

For one thing, it is worth observing that MRE updating is not always so bad. In many successful practical applications, the "constraint" on which to update is of the form $\frac{1}{n}\sum_{i=1}^{n} X_i = t$ for some large n, where X_i is the ith outcome of a random variable X on W. That is, we observe an empirical average of outcomes of X. In such a case, the MRE distribution is "close" (in the appropriate distance measure) to the distribution we arrive at by sophisticated conditioning. That is, if $\Pr'' = \Pr_W(\cdot \mid E(X) = t)$, $\Pr' = \Pr(\cdot \mid X_O =< \frac{1}{n}\sum_{i=1}^{n} X_i = t\rangle)$, and Q^n denotes the n-fold product of a probability distribution Q, then for sufficiently large n, we have that $(\Pr'')^n \approx (\Pr'_W)^n$ (van Campenhout & Cover, 1981; Grünwald, 2001; Skyrms, 1985; Uffink, 1996). Thus, in such cases MRE (almost) coincides with sophisticated conditioning after all. (See (Dawid, 2001) for a discussion of how this result can be reconciled with the results of Section 5.)

But when this special situation does not apply, it is worth asking whether there exists an approach for updating in the naive space that can be easily applied in practical situations, yet leads to *better*, in some formally provable sense, updated distributions than the methods we have considered? A very interesting candidate, often informally applied by human agents, is to simply *ignore* the available extra information. It turns out that there are situations where this update rule behaves better, in a precise sense, than the three methods we have considered. This will be explored in future work.

Another issue that needs further exploration is the subtle distinction between "weak" and "strong" CAR, which was brought to our attention by Manfred Jaeger. The terminology is due to James Robins, who only very recently discovered the distinction. It turns out that the notion of CAR we use in this paper ("weak" CAR) is slightly different from the "strong' version of CAR used by Gill et al. (1997). They view the CAR assumption quite literallly as an assumption about a "coarsening process". Adjusted to our notation, they write (Gill et al., 1997, page 260): "Firstly the random variable X_W of interest is realized; secondly, a *conceptually different* process (usually associated with features of measurement or observational restrictions, rather than the scientific phenomenon under study itself), given the value w taken by X_W, replaces this value by a set U such that $w \in U$." Thus, in their view, the distribution on $\mathcal{R}$ is constructed from a distribution on W and a (conceptually unrelated) set of conditional distributions $\Pr(X_O = \cdot \mid X_W = w)$, one for each $w \in W$. This implies that $\Pr(X_O = \cdot \mid X_W = w)$ is a well-defined number even if $\Pr(X_W = w) = 0$. Gill, van der Laan, and Robins (1997) then *define* the CAR condition as "for all $U \in O$, $\Pr(X_O = U \mid X_W = w)$ is constant in $w \in U$". This is just part (d) of Theorem 3.1, which now has to hold even if $\Pr(X_W = w) = 0$. Thus, the set of distributions satisfying strong CAR is a subset of the set satisfying weak CAR. Robins and Jaeger show that the inclusion can be strict and that this can have substantial consequences. Therefore, our characterizations of CAR in Section 4 apply only to weak CAR, and further research is needed to see the extent to which they also apply to strong CAR.

Our discussion here has focused completely on the probabilistic case. However, these questions also make sense for other representations of uncertainty. Interestingly, Friedman and Halpern (1999) show that that AGM-style belief revision (Alchourrón, Gärdenfors, & Makinson, 1985) can be represented in terms of conditioning using a qualitative representation of uncertainty called a *plausibility measure*; to do this, the plausibility measure must satisfy the analogue of Theorem 3.1(a), so that observations carry no more information than the fact that they are true. No CAR-like condition is given to guarantee that this condition holds for plausibility measures though. It would be interesting to know if there are analogues to CAR for other representations of uncertainty, such as *possibility measures* (Dubois & Prade, 1990) or *belief functions* (Shafer, 1976).

Acknowledgments

A preliminary version of this paper appears in *Uncertainty in Artificial Intelligence, Proceedings of the Eighteenth Conference*, 2002. James Robins and Manfred Jaeger commented on an earlier draft of this paper. We would very much like to thank them for their insightful remarks, which led to several significant improvements in the paper. We also thank the referees of the UAI submission and the *JAIR* submission for their perceptive comments. The first author was supported by a travel grant awarded by the Netherlands Organization for Scientific Research (NWO). The second author was supported in part by NSF under grants IIS-0090145 and CTC-0208535, by ONR under grants N00014-00-1-0341, N00014-01-1-0511, and N00014-02-1-0455, by the DoD Multidisciplinary University Research Initiative (MURI) program administered by the ONR under grants N00014-97-0505 and N00014-01-1-0795, and by a Guggenheim Fellowship, a Fulbright Fellowship, and a grant from NWO. Sabbatical support from CWI and the Hebrew University of Jerusalem is also gratefully acknowledged.

Appendix A. Proofs

In this section, we provide the proofs of all the results in the paper. For convenience, we restate the results here.

Theorem 3.1: *Fix a probability* $\Pr$ *on* $\mathcal{R}$ *and a set* $U \subseteq W$*. The following are equivalent:*

(a) *If* $\Pr(X_O = U) > 0$*, then* $\Pr(X_W = w \mid X_O = U) = \Pr(X_W = w \mid X_W \in U)$ *for all* $w \in U$*.*

(b) *The event* $X_W = w$ *is independent of the event* $X_O = U$ *given* $X_W \in U$*, for all* $w \in U$*.*

(c) $\Pr(X_O = U \mid X_W = w) = \Pr(X_O = U \mid X_W \in U)$ *for all* $w \in U$ *such that* $\Pr(X_W = w) > 0$.

(d) $\Pr(X_O = U \mid X_W = w) = \Pr(X_O = U \mid X_W = w')$ *for all* $w, w' \in U$ *such that* $\Pr(X_W = w) > 0$ *and* $\Pr(X_W = w') > 0$.

Proof: Suppose (a) holds. We want to show that $X_W = w$ and $X_O = U$ are independent, for all $w \in U$. Fix $w \in U$. If $\Pr(X_O = U) = 0$ then the events are trivially independent. So suppose that $\Pr(X_O = U) > 0$. Clearly

$$\Pr(X_W = w \mid X_O = U \cap X_W \in U) = \Pr(X_W = w \mid X_O = U)$$

(since observing U implies that the true world is in U). By part (a),

$$\Pr(X_W = w \mid X_O = U) = \Pr(X_W = w \mid X_W \in U).$$

Thus,

$$\Pr(X_W = w \mid X_U = U \cap X_W \in U) = \Pr(X_W = w \mid X_W \in U),$$

showing that $X_W = w$ is independent of $X_O = U$, given $X_W \in U$.

Next suppose that (b) holds, and $w \in U$ is such that $\Pr(X_W = w) > 0$. From part (b) it is immediate that $\Pr(X_O = U \mid X_W = w \cap X_W \in U) = \Pr(X_O = U \mid X_W \in U)$. Moreover, since $w \in U$, clearly $\Pr(X_O = U \mid X_W = w \cap X_W \in U) = \Pr(X_O = U \mid X_W = w)$. Part (c) now follows.

Clearly (d) follows immediately from (c). Thus, it remains to show that (a) follows from (d). We do this by showing that (d) implies (c) and that (c) implies (a). So suppose that (d) holds. Suppose that $\Pr(X_O = U \mid X_W = w) = a$ for all $w \in U$ such that $\Pr(X_W = w) > 0$. From the definition of conditional probability

$$\begin{array}{rl} & \Pr(X_O = U \mid X_W \in U) \\ = & \sum_{\{w \in U : \Pr(X_W = w) > 0\}} \Pr(X_O = U \cap X_W = w) / \Pr(X_W \in U) \\ = & \sum_{\{w \in U : \Pr(X_W = w) > 0\}} \Pr(X_O = U \mid X_W = w) \Pr(X_W = w) / \Pr(X_W \in U) \\ = & \sum_{\{w \in U : \Pr(X_W = w) > 0\}} a \Pr(X_W = w) / \Pr(X_W \in U) \\ = & a \end{array}$$

Thus, (c) follows from (d).

Finally, to see that (a) follows from (c), suppose that (c) holds. If $w \in U$ is such that $\Pr(X_W = w) = 0$, then (a) is immediate, so suppose that $\Pr(X_W = w) > 0$. Then, using (c) and the fact that $X_O = U \subseteq X_W \in U$, we have that

$$\begin{array}{rl} & \Pr(X_W = w \mid X_O = U) \\ = & \Pr(X_O = U \mid X_W = w) \Pr(X_W = w) / \Pr(X_O = U) \\ = & \Pr(X_O = U \mid X_W \in U) \Pr(X_W = w) / \Pr(X_O = U) \\ = & \Pr(X_O = U \cap X_W \in U) \Pr(X_W = w) / \Pr(X_W \in U) \Pr(X_O = U) \\ = & \Pr(X_O = U) \Pr(X_W = w) / \Pr(X_W \in U) \Pr(X_O = U) \\ = & \Pr(X_W = w) / \Pr(X_W \in U) \\ = & \Pr(X = w \mid X_W \in U), \end{array}$$

as desired. ∎

Proposition 4.1: *The CAR condition holds for all distributions* Pr *on* $\mathcal{R}$ *if and only if* O *consists of pairwise disjoint subsets of* W.

Proof: First suppose that the sets in O are pairwise disjoint. Then for each probability distribution Pr on $\mathcal{R}$, each $U \in O$, and each world $w \in U$ such that $\Pr(X_W = w) > 0$, it must be the case that $\Pr(X_O = U \mid X_W = w) = 1$. Thus, part (d) of Theorem 3.1 applies.

For the converse, suppose that the sets in O are not pairwise disjoint. Then there exist sets $U, U' \in O$ such that both $U - U'$ and $U \cap U'$ are nonempty. Let $w_0 \in U \cap U'$. Clearly there exists a distribution Pr on $\mathcal{R}$ such that $\Pr(X_O = U) > 0$, $\Pr(X_O = U') > 0$, $\Pr(X_W = w_0 \mid X_O = U) = 0, \Pr(X_W = w_0 \mid X_O = U') > 0$. But then $\Pr(X_W = w_0 \mid X_W \in U) > 0$. Thus

$$\Pr(X_W = w_0 \mid X_O = U) \neq \Pr(X_W = w_0 \mid X_W \in U),$$

and the CAR condition (part (a) of Theorem 3.1) is violated. ∎

Lemma 4.3: *Let* $\mathcal{R}$ *be the set of runs over observations* O *and worlds* W, *and let* S *be the CARacterizing matrix for* O *and* W.

(a) *Let* Pr *be any distribution over* $\mathcal{R}$ *and let* S' *be the matrix obtained by deleting from* S *all rows corresponding to an atom* A *with* $\Pr(X_W \in A) = 0$. *Define the vector* $\vec{\gamma} = (\gamma_1, \ldots, \gamma_n)$ *by setting* $\gamma_j = \Pr(X_O = U_j \mid X_W \in U_j)$ *if* $\Pr(X_W \in U_j) > 0$, *and* $\gamma_j = 0$ *otherwise, for* $j = 1, \ldots, n$. *If* Pr *satisfies CAR, then* $S' \cdot \vec{\gamma}^T = \vec{1}^T$.

(b) *Let* S' *be a matrix consisting of a subset of the rows of* S, *and let* $\mathcal{P}_{W,S'}$ *be the set of distributions over* W *with support corresponding to* S'; *i.e.,*

$$\mathcal{P}_{W,S'} = \{P_W \mid P_W(A) > 0 \text{ iff } A \text{ corresponds to a row in } S'\}.$$

If there exists a vector $\vec{\gamma} \geq \vec{0}$ *such that* $S' \cdot \vec{\gamma}^T = \vec{1}^T$, *then, for* all $P_W \in \mathcal{P}_{W,S'}$, *there exists a distribution* Pr *over* $\mathcal{R}$ *with* $\Pr_W = P_W$ *(i.e., the marginal of* Pr *on* W *is* P_W*) such that (a)* Pr *satisfies CAR and (b)* $\Pr(X_O = U_j \mid X_W \in U_j) = \gamma_j$ *for all* j *with* $\Pr(X_W \in U_j) > 0$.

Proof: For part (a), suppose that Pr is a distribution on $\mathcal{R}$ that satisfies CAR. Let k be the number of rows in S', and let $\alpha_i = \Pr(X_W \in A_i)$, for $i = 1, \ldots, k$, where A_i is the atom corresponding to the ith row of S'. Note that $\alpha_i > 0$ for $i = 1, \ldots, k$. Clearly,

$$\sum_{\{j: A_i \subseteq U_j\}} \Pr(X_O = U_j \mid X_W \in A_i) = 1. \tag{6}$$

It easily follows from the CAR condition that

$$\Pr(X_O = U_j \mid X_W \in A_i) = \Pr(X_O = U_j \mid X_W \in U_j)$$

for all $A_i \subseteq U_j$, so (6) is equivalent to

$$\sum_{\{j: A_i \subseteq U_j\}} \Pr(X_O = U_j \mid X_W \in U_j) = 1. \tag{7}$$

(7) implies that $\sum_{\{j: A_i \subseteq U_j\}} \gamma_j = 1$ for $i = 1, \ldots, k$. Let $\vec{s}_i$ be the row in S' corresponding to A_i. Since $\vec{s}_i$ has a 1 as its jth component if $A_i \subseteq U_j$ and a 0 otherwise, it follows that $\vec{s}_i \cdot \vec{\gamma}^T = 1$ and hence $S' \cdot \vec{\gamma}^T = \vec{1}^T$.

For part (b), let k be the number of rows in S', let $\vec{s}_1, \ldots, \vec{s}_k$ be the rows of S', and let $A_1, \ldots, A_k$ be the corresponding atoms. Fix $P_W \in \mathcal{P}_{W,S}$, and set $\alpha_i = P_W(A_i)$ for $i = 1, \ldots, k$. Let Pr be the unique distribution on $\mathcal{R}$ such that

$$\begin{array}{lcl} \Pr(X_W \in A_i) & = & \alpha_i, \text{for } i = 1, \ldots, k, \\ \Pr(X_W \in A) & = & 0 \text{ if } A \in \mathcal{A} - \{A_1, \ldots, A_k\}, \\ \Pr(X_O = U_j \mid X_W \in A_i) & = & \left\{ \begin{array}{ll} \gamma_j & \text{if } A_i \in U_j, \\ 0 & \text{otherwise.} \end{array} \right. \end{array} \tag{8}$$

Note that Pr is indeed a probability distribution on $\mathcal{R}$, since $\sum_{A \in \mathcal{A}} \Pr(X_W \in A) = 1$, $\Pr(X_W \in A_i) > 0$ for $i = 1, \ldots, k$, and, since we are assuming that $S' \cdot \vec{\gamma}^T = \vec{1}^T$,

$$\sum_{j=1}^{n} \Pr(X_O = U_j \mid X_W \in A_i) = \vec{s}_i \cdot \vec{\gamma}^T = 1,$$

for $i = 1, \ldots, k$. Clearly $\Pr_W = P_W$. It remains to show that Pr satisfies CAR and that $\gamma_j = \Pr(X_O = U_j \mid X_W \in U_j)$. Given $j \in \{1, \ldots, n\}$, suppose that there exist atoms $A_i, A_{i'}$ corresponding to rows $\vec{s}_i$ and $\vec{s}_{i'}$ of S' such that $A_i, A_{i'} \in U_j$. Then

$$\Pr(X_O = U_j | X_W \in A_i) = \Pr(X_O = U_j | X_W \in A_{i'}) = \gamma_j.$$

It now follows by Theorem 3.1(c) that Pr satisfies the CAR condition for $U_1, \ldots, U_n$. Moreover, Theorem 3.1(d), it must be the case that $\Pr(X_O = U_j \mid X_W \in U_j) = \gamma_j$. ∎

The proof of Theorem 4.4 builds on Lemma 4.3 and the following proposition, which shows that the condition of part (b) of Theorem 4.4 is actually stronger than the condition of part (a). It is therefore not surprising that it leads to a stronger conclusion.

Proposition A.1: *If there exists a subset R of rows of S that is linearly dependent but not affinely dependent, then for all $\mathcal{R}$-atoms A corresponding to a row in R and all $j^* \in \{1, \ldots, n\}$, if $A \subseteq U_{j^*}$, there exists a vector $\vec{u}$ that is an affine combination of the rows in R such that $u_j \geq 0$ for all $j \in \{1, \ldots, n\}$ and $u_{j^*} > 0$.*

Proof: Suppose that there exists a subset R of rows of S that is linearly dependent but not affinely dependent. Without loss of generality, let $\vec{v}_1, \ldots, \vec{v}_k$ be the rows in R. There exist $\lambda_1, \ldots, \lambda_k$ such that $\kappa = \sum_{i=1}^k \lambda_i \neq 0$ and $\sum_{i=1}^k \lambda_i \vec{v}_i = 0$. We first show that in fact every row $\vec{v}$ in R is an affine combination of the other rows. Fix some $j \in \{1, \ldots, k\}$. Let $\mu_j = (\lambda_j - \sum_{i=1}^k \lambda_i) = -\sum_{i \neq j} \lambda_i$ and let $\mu_i = \lambda_i$ for $i \neq j$. Then $\sum_{i=1}^k \mu_i = 0$ and

$$\sum_{i=1}^{k} \mu_i \vec{v}_i = \sum_{i=1}^{k} \lambda_i \vec{v}_i - \sum_{i=1}^{k} \lambda_i \vec{v}_j = -\kappa \vec{v}_j.$$

For $i = 1, \ldots, k$, let $\mu_i' = -\mu_i/\kappa$. Then $\sum_{i=1}^k \mu_i' = 0$ and $\sum_{i=1}^k \mu_i' \vec{v}_i = \vec{v}_j$. Now if $A_i \subseteq U_{j^*}$ for some $i = 1, \ldots, k$ and some $j^* = 1, \ldots, n$, then $\vec{v}_i$ has a 1 as its j^*th component. Also, $\vec{v}_i$ is an affine combination of the rows of R with no negative components, so $\vec{v}_i$ is the desired vector. ∎

Theorem 4.4: *Let $\mathcal{R}$ be a set of runs over observations $O = \{U_1, \ldots, U_n\}$ and worlds W, and let S be the CARacterizing matrix for O and W.*

(a) Suppose that there exists a subset R of the rows in S and a vector $\vec{u} = (u_1, \ldots, u_n)$ that is an affine combination of the rows of R such that $u_j \geq 0$ for all $j \in \{1, \ldots, n\}$ and $u_{j^*} > 0$ for some $j^* \in \{1, \ldots, n\}$. Then there is no distribution Pr *on $\mathcal{R}$ that satisfies CAR such that* $\Pr(X_O = U_{j^*}) > 0$ *and* $\Pr(X_W \in A) > 0$ *for each $\mathcal{R}$-atom A corresponding to a row in R.*

(b) If there exists a subset R of the rows of S that is linearly dependent but not affinely dependent, then there is no distribution Pr *on $\mathcal{R}$ that satisfies CAR such that* $\Pr(X_W \in A) > 0$ *for each $\mathcal{R}$-atom A corresponding to a row in R.*

(c) Given a set R consisting of n linearly independent rows of S and a distribution P_W on W such that $P_W(A) > 0$ for all A corresponding to a row in R, there is a unique distribution P_O on O such that if Pr *is a distribution on $\mathcal{R}$ satisfying CAR and* $\Pr(X_W \in A) = P_W(A)$ *for each atom A corresponding to a row in R, then* $\Pr(X_O = U) = P_O(U)$.

Proof: For part (a), suppose that R consists of $\vec{v}_1, \ldots, \vec{v}_k$, corresponding to atoms $A_1, \ldots, A_k$. By assumption, there exist coefficients $\lambda_1, \ldots, \lambda_k$ such that $\sum_{i=1}^k \lambda_i = 0$, and a vector $\vec{u} = \sum_{i=1}^k \lambda_i \vec{v}_i$ such that every component of $\vec{u}$ is nonnegative. Suppose, by way of contradiction, that Pr satisfies CAR and that $\alpha_i = \Pr(X_W \in A_i) > 0$ for $i \in \{1, \ldots, k\}$. By Lemma 4.3(a), we have

$$\vec{u} \cdot \vec{\gamma} = \left(\sum_{i=1}^k \lambda_i \vec{v}_i \right) \cdot \vec{\gamma} = \sum_{i=1}^k \lambda_i (\vec{v}_i \cdot \vec{\gamma}) = \sum_{i=1}^k \lambda_i = 0, \tag{9}$$

where $\vec{\gamma}$ is defined as in Lemma 4.3. For $j = 1, \ldots, n$, if $\Pr(X_O = U_j) > 0$ then $\Pr(X_O = U_j \cap X_W \in U_j) = \Pr(X_O = U_j) > 0$ and $\Pr(X_W \in U_j) > 0$, so $\gamma_j > 0$. By assumption, all the components of $\vec{u}$ and $\vec{\gamma}$ are nonnegative. Therefore, if there exists j^* such that $\Pr(X_O = U_{j^*}) > 0$ and $u_{j^*} > 0$, then $\vec{u} \cdot \vec{\gamma} > 0$. This contradicts (9), and part (a) is proved.

For part (b), suppose that there exists a subset R of rows of S that is linearly dependent but not affinely dependent. Suppose, by way of contradiction, that Pr satisfies CAR and that $\Pr(X_W \in A) > 0$ for all atoms A corresponding to a row in R. Pick an atom A^* corresponding to such a row. By Proposition A.1 and Theorem 4.4(a), we have that $\Pr(X_O = U_{j^*}) = 0$ for all j^* such that $A^* \in U_{j^*}$. But then $\Pr(X_W \in A^*) = 0$, and we have arrived at a contradiction.

For part (c), suppose that R consists of the rows $\vec{v}_1, \ldots, \vec{v}_n$. Let S' be the $n \times n$ submatrix of S consisting of the rows of R. Since these rows are linearly independent, a standard result of linear algebra says that S' is invertible. Let Pr be a distribution on $\mathcal{R}$ satisfying CAR. By Lemma 4.3(a), $S'\vec{\gamma} = \vec{1}^T$. Thus, $\vec{\gamma} = (S')^{-1}\vec{1}$. For $j = 1, \ldots, n$ we must have $\gamma_j = \beta_j / \Pr(X_W \in U_j)$, where $\beta_j = \Pr(X_O = U_j)$. Given $\Pr_W(A)$ for each atom A, we can clearly solve for the β_j's. ∎

Theorem 4.9: *Given a set $\mathcal{R}$ of runs over a set W of worlds and a set O of observations,* Pr *is a distribution on $\mathcal{R}$ that satisfies CAR iff there is a setting of the parameters in* CARgen* *such that, for all $w \in W$ and $U \in O$,* $\Pr(\{r : X_W(r) = w, X_O(r) = U\})$ *is the probability that* CARgen* *returns (w, U).*

Proof: First we show that if Pr is a probability on $\mathcal{R}$ such that, for some setting of the parameters of CARgen*, $\Pr(\{r : X_W(r) = w, X_O(r) = U\})$ is the probability that CARgen* returns (w, U), then Pr satisfies CAR. By Theorem 3.1, it suffices to show that, for each set $U \in O$ and worlds $w_1, w_2 \in U$ such that $\Pr(X_W = w_1) > 0$ and $\Pr(X_W = w_2) > 0$, we have $\Pr(X_O = U \mid X_W = w_1) = \Pr(X_O = U \mid X_W = w_2)$. So suppose that $w_1, w_2 \in U$, $\Pr(X_W = w_1) > 0$, and $\Pr(X_W = w_2) > 0$. Let $\alpha_U = \sum_{\{\Pi \in \mathcal{P} : U \in \Pi\}} P_{\mathcal{P}}(\Pi)(1 - q_{U|\Pi})$. Intuitively, α_U is the probability that the algorithm terminates immediately at step 2.3 with (w, U) conditional on some $w \in U$ being chosen at step 2.1. Notice for future reference that, for all w,

$$\sum_{\{U : w \in U\}} \alpha_U = \sum_{\{(U,\Pi) : U \in \Pi, w \in U\}} P_{\mathcal{P}}(\Pi)(1 - q_{U|\Pi}) = 1 - q, \tag{10}$$

where q is defined by (4). As explained in the main text, for both $i = 1, 2$, q is the probability that the algorithm does not terminate at step 2.3 given that w_i is chosen in step 2.1. It easily follows that the probability that (w_i, U) is output at step 2.3 is

$$P_W(w_i)\alpha_U(1 + q + q^2 + \cdots) = P_W(w_i)\alpha_U/(1-q).$$

Thus, $\Pr(X_W = w_i \cap X_O = U) = P_W(w_i)\alpha_U/(1-q)$. Using (10), we have that

$$\Pr(X_W = w_i) = \sum_{\{U: w_i \in U\}} \Pr(X_W = w_i \cap X_O = U) = \frac{P_W(w_i)}{1-q} \sum_{\{U: w_i \in U\}} \alpha_U = P_W(w_i).$$

Finally, we have that $\Pr(X_O = U \mid X_W = w_i) = \alpha_U/(1-q)$, for $i = 1, 2$. Thus, Pr satisfies the CAR condition.

For the converse, suppose that Pr satisfies the CAR condition. Let $O = \{U_1, \ldots, U_n\}$. We choose the parameters for CARGEN* as follows. Set $P_W(w) = \Pr(X_W = w)$ and let $\beta_i = \Pr(X_O = U_i)$. Without loss of generality, we assume that $\beta_i > 0$ (otherwise, take O' to consist of those sets that are observed with positive probability, and do the proof using O').

For $i = 1, \ldots, n$, let $\Pi_i = \{U_i, \overline{U_i}\}$. Set $P_{\mathcal{P}}(\Pi_i) = \Pr(X_O = U_i) = \beta_i$ and $q_{\overline{U}_i|\Pi_i} = 1$. (Thus, the set $\overline{U}_i$ is always rejected, unless $\overline{U}_i = U_j$.) Since $\Pr(X_W \in U_j) \geq \Pr(X_O = U_j) > 0$ by assumption, it must be the case that $\epsilon = \min_{j=1}^n \Pr(X_W \in U_j) > 0$. Now set $q_{U_i|\Pi_i} = 1 - \epsilon/\Pr(X_W \in U_i)$.

We first show that, with these parameter settings, we can choose q such that constraint (4) is satisfied. Let $q_w = \sum_{\{U,\Pi:\, w \in U,\, U \in \Pi\}} P_{\mathcal{P}}(\Pi) q_{U|\Pi}$. For each $w \in W$ such that $P_W(w) > 0$, we have

$$\begin{array}{cl} & q_w \\ = & \sum_{\{U,\Pi:\, w\in U,\, U\in\Pi\}} P_{\mathcal{P}}(\Pi) q_{U|\Pi} \\ = & \sum_{i=1}^n \sum_{\{U:\, w\in U,\, U\in\Pi_i\}} P_{\mathcal{P}}(\Pi_i) q_{U|\Pi_i} \\ = & \sum_{\{i: w\in U_i\}} P_{\mathcal{P}}(\Pi_i) q_{U_i|\Pi_i} + \sum_{\{i: w\in \overline{U}_i\}} P_{\mathcal{P}}(\Pi_i) q_{\overline{U}_i|\Pi_i}. \end{array}$$

The last equality follows because $\Pi = \{U_i, \overline{U}_i\}$. Thus, for a fixed i, $\sum_{\{U:\, w\in U,\, U\in\Pi_i\}} P_{\mathcal{P}}(\Pi_i) q_{U|\Pi_i}$ is either $P_{\mathcal{P}}(\Pi_i) q_{U_i|\Pi_i}$ if $w \in U_i$, or $P_{\mathcal{P}}(\Pi_i) q_{\overline{U}_i|\Pi_i}$ if $w \in \overline{U}_i$. It follows that

$$\begin{array}{cll} & q_w & \\ = & \sum_{\{i: w\in U_i\}} \beta_i (1 - \epsilon/\Pr(X_W \in U_i)) + \sum_{\{i: w\notin U_i\}} \beta_i \cdot 1 & \\ = & \sum_{\{i: w\in U_i\}} \Pr(X_O = U_i)(1 - \epsilon/\Pr(X_W \in U_i)) + \sum_{\{i: w\notin U_i\}} \Pr(X_O = U_i) & \\ = & \sum_{i=1}^n \Pr(X_O = U_i) - \epsilon \sum_{\{i: w\in U_i\}} \Pr(X_O = U_i \mid X_W \in U_i) & \\ = & 1 - \epsilon \sum_{\{i: w\in U_i\}} \Pr(X_O = U_i \mid X_W = w) & \text{[since Pr satisfies CAR]} \\ = & 1 - \epsilon. & \end{array}$$

Thus, $q_w = q_{w'}$ if $P_W(w), P_W(w') > 0$, so these parameter settings are appropriate for CARGEN* (taking $q = q_w$ for any w such that $P_W(w) > 0$). Moreover, $\epsilon = 1 - q$.

We now show that, with these parameter settings, $\Pr(X_W = w \cap X_O = U)$ is the probability that CARGEN* halts with (w, U), for all $w \in W$ and $U \in O$. Clearly if $\Pr(X_W = w) = 0$, this is true, since then $\Pr(X_W = w \cap X_O = U) = 0$, and the probability that CARGEN* halts with output (w, U) is at most $P_W(w) = \Pr(X_W = w) = 0$. So suppose that $\Pr(X_W = w) > 0$. Then it suffices to show that $\Pr(X_O = U_i \mid X_W = w)$ is the probability that (w, U_i) is output, given that w is chosen at the first step. But the argument of the first half of the proof shows that this probability is just $\frac{\alpha_{U_i}}{1-q}$. But

$$\begin{array}{rll} & \frac{\alpha_{U_i}}{1-q} & \\ = & \frac{\alpha_{U_i}}{\epsilon} & [\text{since } \epsilon = 1-q] \\ = & \frac{\sum_{\{\Pi \in \mathcal{P}: U_i \in \Pi\}} P_{\mathcal{P}}(\Pi)(1-q_{U_i \mid \Pi})}{\epsilon} & \\ = & \frac{\beta_i(\epsilon / \Pr(X_W \in U_i))}{\epsilon} & \\ = & \Pr(X_O = U_i) / \Pr(X_W \in U_i) & \\ = & \Pr(X_O = U_i \mid X_W = w) & [\text{since Pr satisfies CAR}], \end{array}$$

as desired. ∎

Theorem 5.1: *Fix a probability* $\Pr$ *on* $\mathcal{R}$, *a partition* $\{U_1, \ldots, U_n\}$ *of* W, *and probabilities* $\alpha_1, \ldots, \alpha_n$ *such that* $\alpha_1 + \cdots + \alpha_n = 1$. *Let* C *be the observation* $\alpha_1 U_1; \ldots; \alpha_n U_n$. *Fix some* $i \in \{1, \ldots, n\}$. *Then the following are equivalent:*

(a) If $\Pr(X_O = C) > 0$, *then* $\Pr(X_W = w \mid X_O = C) = \Pr_W(w \mid \alpha_1 U_1; \ldots; \alpha_n U_n)$ *for all* $w \in U_i$.

(b) $\Pr(X_O = C \mid X_W = w) = \Pr(X_O = C \mid X_W \in U_i)$ *for all* $w \in U_i$ *such that* $\Pr(X_W = w) > 0$.

Proof: The proof is similar in spirit to that of Theorem 3.1. Suppose that (a) holds, $w \in U_i$, and $\Pr(X_W = w) > 0$. Then

$$\begin{array}{rl} & \Pr(X_O = C \mid X_W = w) \\ = & \Pr(X_W = w \mid X_O = C) \Pr(X_O = C) / \Pr(X_W = w) \\ = & \Pr_W(w \mid \alpha_1 U_1; \ldots; \alpha_n U_n) \Pr(X_O = C) / \Pr(X_W = w) \\ = & \alpha_i \Pr_W(w \mid U_i) \Pr(X_O = C) / \Pr_W(w) \\ = & \alpha_i \Pr(X_O = C) / \Pr_W(U_i) \end{array}$$

Similarly,

$$\begin{array}{rl} & \Pr(X_O = C \mid X_W \in U_i) \\ = & \Pr(X_W \in U_i \mid X_O = C) \Pr(X_O = C) / \Pr(X_W \in U_i) \\ = & \sum_{w' \in U_i} \Pr_W(w' \mid \alpha_1 U_1; \ldots; \alpha_n U_n) \Pr(X_O = C) / \Pr(X_W \in U_i) \\ = & \sum_{w' \in U_i} \alpha_i \Pr_W(w' \mid U_i) \Pr(X_O = C) / \Pr(X_W \in U_i) \\ = & \alpha_i \Pr(X_O = C) / \Pr_W(U_i) \end{array}$$

Thus, $\Pr(X_O = C \mid X_W = w) = \Pr(X_O = C \mid X_W \in U_i)$ for all $w \in U_i$ such that $\Pr(X_W = w) > 0$.

For the converse, suppose that (b) holds and $\Pr(X_O = C) > 0$. Given $w \in U_i$, if $\Pr(X_W = w) = 0$, then (a) trivially holds, so suppose that $\Pr(r(X_W = w) > 0$. Suppose that $w \in U_i$. Clearly $\Pr(w \mid \alpha_1 U_1; \ldots; \alpha_n U_n) = \alpha_i \Pr_W(w \mid U_i)$. Now, using (b), we have that

$$\begin{array}{rl} & \Pr(X_W = w \mid X_O = C) \\ = & \Pr(X_O = C \mid X_W = w) \Pr(X_W = w) / \Pr(X_O = C) \\ = & \Pr(X_O = C \mid X_W \in U_i) \Pr(X_W = w) / \Pr(X_O = C) \\ = & \Pr(X_W \in U_i \mid X_O = C) \Pr(X_W = w) / \Pr(X_W \in U_i) \\ = & \alpha_i \Pr_W(w \mid U_i) \quad [\text{using (5)}]. \end{array}$$

Thus, (a) holds. ∎

Proposition 5.2: *Consider a partition $\{U_1, \ldots, U_n\}$ of W and a set of $k > 1$ observations $O = \{C_1, \ldots, C_k\}$ with $C_i = \alpha_{i1}U_1; \ldots; \alpha_{in}U_n$ such that all $\alpha_{ij} > 0$. For every distribution P_O on O with $P_O(C_i) > 0$ for all $i \in \{1, \ldots, k\}$, there exists a distribution $\Pr$ on $\mathcal{R}$ such that $P_O = \Pr_O$ (i.e. P_O is the marginal of $\Pr$ on O) and $\Pr$ satisfies the generalized CAR condition (part (b) of Theorem 5.1) for $U_1, \ldots, U_n$.*

Proof: Given a set W of worlds, a set $O = \{C_1, \ldots, C_k\}$ of observations with distribution P_O satisfying $P_O(C_i) > 0$ for $i \in \{1, \ldots, k\}$, and arbitrary distributions $\Pr_j$ on U_j, $j = 1, \ldots, n$, we explicitly construct a prior $\Pr$ on $\mathcal{R}$ that satisfies CAR such that $P_O = \Pr_O$, where $\Pr_O$ is the marginal of $\Pr$ on O and $\Pr_j = \Pr_W(\cdot \mid U_j)$.

Given $w \in U_j$, define

$$\Pr(\{r \in \mathcal{R} : X_O(r) = C_i, X_W(r) = w\}) = P_O(C_i)\alpha_{ij}\Pr_j(w).$$

(How the probability is split up over all the runs r such that $X_O(r) = C_i$ and $X_W(r) = w$ is irrelevant.) It remains to check that $\Pr$ is a distribution on $\mathcal{R}$ and that it satisfies all the requirements. It is easy to check that

$$\Pr(X_O = C_i) = \sum_{j=1}^{n} \sum_{w \in U_j} P_O(C_i)\alpha_{ij}\Pr_j(w) = P_O(C_i).$$

It follows that $\sum_{i=1}^{k} \Pr(X_O = C_i) = 1$, showing that $\Pr$ is a probability measure and P_O is the marginal of $\Pr$ on O. If $w \in U_j$, then

$$\begin{array}{rl} & \Pr_W(w \mid U_j) = \Pr_W(w)/\Pr_W(U_j) \\ = & \frac{\sum_{i=1}^{k} \Pr_O(C_i)\alpha_{ij}\Pr_j(w)}{\sum_{w' \in U_j} \sum_{i=1}^{k} \Pr_O(C_i)\alpha_{ij}\Pr_j(w')} \\ = & \frac{\Pr(w)\sum_{i=1}^{k} \Pr_O(C_i)\alpha_{ij}}{(\sum_{w' \in U_j} \Pr_j(w'))\sum_{i=1}^{k} \Pr_O(C_i)\alpha_{ij}} \\ = & \Pr_j(w). \end{array}$$

Finally, note that, for $j \in \{1, \ldots, n\}$, for all $w \in U_j$ such that $\Pr(X_W = w) > 0$, we have that

$$\begin{array}{rl} & \Pr(X_O = C_i \mid X_W = w) \\ = & \frac{\Pr_O(C_i)\alpha_{ij}\Pr_j(w)}{\sum_{i=1}^{k} \Pr_O(C_i)\alpha_{ij}\Pr_j(w)} \\ = & \frac{\Pr_O(C_i)\alpha_{ij}}{\sum_{i=1}^{k} \Pr_O(C_i)\alpha_{ij}} \\ = & \frac{\Pr_O(C_i)\alpha_{ij}\Pr(X_W \in U_j)}{\sum_{i=1}^{k} \Pr_O(C_i)\alpha_{ij}\Pr(X_W \in U_j)} \\ = & \frac{\Pr(X_O = C_i \cap X_W \in U_j)}{\Pr(X_W \in U_j)} \\ = & \Pr(X_O = C_i \mid X_W \in U_j) \end{array}$$

so the generalized CAR condition holds for $\{U_1, \ldots, U_n\}$. ∎

To prove Theorem 5.3, we first need some background on minimum relative entropy distributions. Fix some space W and let $U_1, \ldots, U_n$ be subsets of W. Let Δ be the set of

$(\alpha_1, \ldots, \alpha_n)$ for which there exists some distribution P_W with $P_W(U_i) = \alpha_i$ for $i = 1, \ldots, n$ and $P_W(w) > 0$ for all $w \in W$. Now let P_W be a distribution with $P_W(w) > 0$ for all $w \in W$. Given a vector $\vec{\beta} = (\beta_1, \ldots, \beta_n) \in \mathbf{R}^n$, let

$$P_W^{\vec{\beta}}(w) = \frac{1}{Z} e^{\beta_1 \mathbf{1}_{U_1}(w) + \ldots + \beta_n \mathbf{1}_{U_n}(w)} P_W(w),$$

where $\mathbf{1}_U$ is the indicator function, i.e. $\mathbf{1}_U(w) = 1$ if $w \in U$ and 0 otherwise, and $Z = \sum_{w \in W} e^{\beta_1 \mathbf{1}_{U_1}(w) + \ldots + \beta_n \mathbf{1}_{U_n}(w)} P_W(w)$ is a normalization factor. Let $\alpha_i = P_W^{\vec{\beta}}(U_i)$ for $i = 1, \ldots, n$. By (Csiszár, 1975, Theorems 2.1 and 3.1), it follows that

$$P_W(\cdot \mid \alpha_1 U_1; \ldots; \alpha_n U_n) = P_W^{\vec{\beta}}; \tag{11}$$

Moreover, for each vector $(\alpha_1, \ldots, \alpha_n) \in \Delta$, there is a vector $\vec{\beta} = (\beta_1, \ldots, \beta_n) \in \mathbf{R}^n$ such that (11) holds. (For an informal and easy derivation of (11), see (Cover & Thomas, 1991, Chapter 9).)

Lemma A.2: *Let $C = \alpha_1 U_1; \ldots; \alpha_n U_n$ for some $(\alpha_1, \ldots, \alpha_n) \in \Delta$. Let $(\beta_1, \ldots, \beta_n)$ be a vector such that (11) holds for $\alpha_1, \ldots, \alpha_n$. If $\beta_i = 0$ for some $i \in \{1, \ldots, n\}$, then*

$$P_W(U_i \mid \alpha_1 U_1; \ldots; \alpha_{i-1} U_{i-1}; \alpha_{i+1} U_{i+1}; \ldots; \alpha_n U_n) = \alpha_i.$$

Proof: Without loss of generality, assume that $\beta_1 = 0$. Taking $\alpha_i' = P_W^{\vec{\beta}}(U_i)$ for $i = 2, \ldots, n$, it follows from (11) that

$$P_W(w \mid \alpha_2' U_2; \ldots; \alpha_n' U_n') = \frac{1}{Z} e^{\beta_2 \mathbf{1}_{U_2} + \ldots + \beta_n \mathbf{1}_{U_n}} P_W(w),$$

so that

$$P_W(\cdot \mid \alpha_1 U_1; \ldots; \alpha_n U_n) = P_W(\cdot \mid \alpha_2' U_2; \ldots; \alpha_n' U_n).$$

Since $P_W(U_i \mid \alpha_2' U_2; \ldots; \alpha_n' U_n) = \alpha_i'$ and $P_W(U_i \mid \alpha_1 U_1; \ldots \alpha_n U_n) = \alpha_i$ for $i = 2, \ldots, n$, we have that $\alpha_i = \alpha_i'$ for $i = 2, \ldots, n$. Thus, $P_W(\cdot \mid \alpha_2 U_2; \ldots; \alpha_n U_n) = P_W(\cdot \mid \alpha_1 U_1; \ldots; \alpha_n U_n)$ and, in particular,

$$\alpha_1 = P_W(U_1 \mid \alpha_1 U_1; \ldots; \alpha_n U_n) = P_W(U_1 \mid \alpha_2 U_2; \ldots; \alpha_n U_n).$$

∎

Theorem 5.3: *Given a set $\mathcal{R}$ of runs and a set $O = \{C_1, C_2\}$ of observations, where $C_i = \alpha_{i1} U_1; \alpha_{i2} U_2$, for $i = 1, 2$, let $\Pr$ be a distribution on $\mathcal{R}$ such that $\Pr(X_O = C_1)$, $\Pr(X_O = C_2) > 0$, and $\Pr_W(w) = \Pr(X_W = w) > 0$ for all $w \in W$. Let $\Pr^i = \Pr(\cdot \mid X_O = C_i)$, and let $\Pr^i_W$ be the marginal of $\Pr^i$ on W. If either C_1 or C_2 is not Jeffrey-like, then we cannot have $\Pr^i_W = \Pr_W(\cdot \mid C_i)$, for both $i = 1, 2$.*

Proof: Let $V_1 = U_1 - U_2, V_2 = U_2 - U_1, V_3 = U_1 \cap U_2$, and $V_4 = W - (U_1 \cup U_2)$. Since V_1, V_2, V_3, V_4 are all assumed to be nonempty, we have $\Delta = (0,1)^2$, where Δ is defined as above, that is, Δ is the set (α_1, α_2) such that there exists a distribution P_W with $P_W(U_1) = \alpha_1, P_W(U_2) = \alpha_2$, $P_W(w) > 0$ for all $w \in W$. If $\Pr^i_W = \Pr_W(\cdot \mid C_i)$ for $i = 1, 2$, then

$$\lambda \Pr_W(\cdot | C_1) + (1 - \lambda) \Pr_W(\cdot \mid C_2) = \Pr_W, \tag{12}$$

where $\lambda = \Pr(X_O = C_1)$. We prove the theorem by showing that (12) cannot hold if either C_1 or C_2 is not Jeffrey-like. Since we have assumed that $(\alpha_{i1}, \alpha_{i2}) \in (0,1)^2 = \Delta$ for $i = 1,2$, we can apply (11) to C_i for $i = 1,2$. Thus, there are vectors $(\beta_{i1}, \beta_{i2}) \in \mathbf{R}^2$ for $i = 1,2$ such that, for all $w \in W$,

$$\Pr_W(w|C_i) = \frac{1}{Z_i} e^{\beta_{i1}\mathbf{1}_{U_1} + \beta_{i2}\mathbf{1}_{U_2}} \Pr_W(w). \tag{13}$$

(13) implies $\Pr_W(V_1|C_i) = Z_i^{-1} e^{\beta_{i1}} \Pr_W(V_1)$, $\Pr_W(V_2|C_i) = Z_i^{-1} e^{\beta_{i2}} \Pr_W(V_2)$, $\Pr_W(V_3|C_i) = Z_i^{-1} e^{\beta_{i1}+\beta_{i2}} \Pr_W(V_3)$, $\Pr_W(V_4|C_i) = Z_i^{-1} \Pr_W(V_4)$. Plugging this into (12), we obtain the following four equations:

$$\begin{aligned}
\Pr_W(V_1) &= \lambda \frac{e^{\beta_{11}}}{Z_1} \Pr_W(V_1) + (1-\lambda) \frac{e^{\beta_{21}}}{Z_2} \Pr_W(V_1) \\
\Pr_W(V_2) &= \lambda \frac{e^{\beta_{12}}}{Z_1} \Pr_W(V_2) + (1-\lambda) \frac{e^{\beta_{22}}}{Z_2} \Pr_W(V_2) \\
\Pr_W(V_3) &= \lambda \frac{e^{\beta_{11}+\beta_{21}}}{Z_1} \Pr_W(V_3) + (1-\lambda) \frac{e^{\beta_{21}+\beta_{22}}}{Z_2} \Pr_W(V_3) \\
\Pr_W(V_4) &= \lambda \frac{1}{Z_1} \Pr_W(V_4) + (1-\lambda) \frac{1}{Z_2} \Pr_W(V_4).
\end{aligned} \tag{14}$$

Since we have assumed that $\Pr(w) > 0$ for all $w \in W$, it must be the case that $\Pr_W(V_i) > 0$, for $i = 1, \ldots, 4$. Thus, $\Pr(V_i)$ factors out of the ith equation above. By the change of variables $\mu = \lambda/Z_1$, $1 - \mu = (1-\lambda)/Z_2$, $\epsilon_{ij} = e^{\beta_{ij}} - 1$ and some rewriting, we see that (14) is equivalent to

$$\begin{aligned}
0 &= \mu\epsilon_{11} + (1-\mu)\epsilon_{21} \\
0 &= \mu\epsilon_{12} + (1-\mu)\epsilon_{22} \\
0 &= \mu(\epsilon_{11} + \epsilon_{12} + \epsilon_{11}\epsilon_{12}) + (1-\mu)(\epsilon_{21} + \epsilon_{22} + \epsilon_{21}\epsilon_{22}).
\end{aligned} \tag{15}$$

If, for some i, both ϵ_{i1} and ϵ_{i2} are nonzero, then the three equations of (15) have no solutions for $\mu \in (0,1)$. Equivalently, if for some i, both β_{i1} and β_{i2} are nonzero, then the four equations of (14) have no solutions for $\lambda \in (0,1)$. So it only remains to show that for some i, both β_{i1} and β_{i2} are nonzero. To see this, note that by assumption for some i, C_i is not Jeffrey-like. But then it follows from Lemma A.2 above that both β_{i1} and β_{i2} are nonzero. Thus, the theorem is proved. ∎

References

Alchourrón, C. E., Gärdenfors, P., & Makinson, D. (1985). On the logic of theory change: partial meet functions for contraction and revision. *Journal of Symbolic Logic*, *50*, 510–530.

Bar-Hillel, M., & Falk, R. (1982). Some teasers concerning conditional probabilities. *Cognition*, *11*, 109–122.

Cover, T. M., & Thomas, J. A. (1991). *Elements of Information Theory*. Wiley, New York.

Csiszár, I. (1975). *I*-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, *3*(1), 146–158.

Csiszár, I. (1991). Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems. *Annals of Statistics*, *19*(4), 2032–2066.

Dawid, A. P. (2001). A note on maximum entropy and Bayesian conditioning.. Unpublished manuscript.

Dawid, A. P., & Dickey, J. M. (1977). Likelihood and Bayesian inference from selectively reported data. *Journal of the American Statistical Association*, *72*(360), 845–850.

Diaconis, P., & Zabell, S. L. (1986). Some alternatives to Bayes's rule. In Grofman, B., & Owen, G. (Eds.), *Proc. Second University of California, Irvine, Conference on Political Economy*, pp. 25–38.

Dubois, D., & Prade, H. (1990). An introduction to possibilistic and fuzzy logics. In Shafer, G., & Pearl, J. (Eds.), *Readings in Uncertain Reasoning*, pp. 742–761. Morgan Kaufmann, San Francisco.

Freund, J. E. (1965). Puzzle or paradox?. *American Statistician*, *19*(4), 29–44.

Friedman, K., & Shimony, A. (1971). Jaynes' maximum entropy prescription and probability theory. *Journal of Statistical Physics*, *9*, 265–269.

Friedman, N., & Halpern, J. Y. (1997). Modeling belief in dynamic systems. Part I: foundations. *Artificial Intelligence*, *95*(2), 257–316.

Friedman, N., & Halpern, J. Y. (1999). Modeling belief in dynamic systems. Part II: revision and update. *Journal of A.I. Research*, *10*, 117–167.

Gardner, M. (1961). *Second Scientific American Book of Mathematical Puzzles and Diversions*. Simon & Schuster, New York.

Gardner, M. (1982). *Aha, Gotcha!* WH Freeman & Co.

Gill, R. D., van der Laan, M., & Robins, J. (1997). Coarsening at random: Characterisations, conjectures and counter-examples. In *Proc. First Seattle Conference on Biostatistics*, pp. 255–294.

Grove, A. J., & Halpern, J. Y. (1997). Probability update: conditioning vs. cross-entropy. In *Proc. Thirteenth Conference on Uncertainty in Artificial Intelligence (UAI '97)*, pp. 208–214.

Grünwald, P. (2001). Strong entropy concentration, game theory and algorithmic randomness. In *Proc. Fourteenth Annual Conference on Computational Learning Theory*, pp. 320–336.

Halpern, J. Y., & Fagin, R. (1989). Modelling knowledge and action in distributed systems. *Distributed Computing*, *3*(4), 159–179.

Halpern, J. Y., & Tuttle, M. R. (1993). Knowledge, probability, and adversaries. *Journal of the ACM*, *40*(4), 917–962.

Heitjan, D., & Rubin, D. (1991). Ignorability and coarse data. *Annals of Statistics*, *19*, 2244–2253.

Hunter, D. (1989). Causality and maximum entropy updating. *International Journal of Approximate Reasoning*, *3*(1), 379–406.

Jeffrey, R. C. (1968). Probable knowledge. In Lakatos, I. (Ed.), *International Colloquium in the Philosophy of Science: The Problem of Inductive Logic*, pp. 157–185. North-Holland, Amsterdam.

Kleinbaum, D. (1999). *Survival Analysis: A Self-Learning Text*. Statistics in the Health Sciences. Springer-Verlag, New York.

Kullback, S. (1959). *Information Theory and Statistics*. Wiley.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, *22*, 76–86.

Mosteller, F. (1965). *Fifty Challenging Problems in Probability with Solutions*. Addison-Wesley, Reading, Mass.

Nielsen, S. (1998). *Coarsening at Random and Simulated EM Algorithms*. Ph.D. thesis, Department of Theoretical Statistics, University of Copenhagen.

Robins, J., Rotnitzky, A., & Scharfstein, D. (1999). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In Hallor, M. E., & Berry, D. (Eds.), *Statistical Models in Epidemiology: The Environment and Clinical Trials*, IMA Volume 116, pp. 1–92. Springer-Verlag.

Rubin, D. (1976). Inference and missing data. *Biometrika*, *63*, 581–592.

Scharfstein, D. O., Daniels, M., & Robins, J. M. (2002). Incorporating prior beliefs about selection bias into the analysis of randomized trials with missing outcomes. *Biometrika*. To appear.

Seidenfeld, T. (1986). Entropy and uncertainty. *Philosophy of Science*, *53*, 467–491.

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

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

Shore, J. E., & Johnson, R. W. (1980). Axiomatic derivation of the principle of maximum entropy and the principle of minimimum cross-entropy. *IEEE Transactions on Information Theory*, *IT-26*(1), 26–37.

Skyrms, B. (1985). Maximum entropy inference as a special case of conditionalization. *Synthese*, *63*, 55–74.

Uffink, J. (1996). The constraint rule of the maximum entropy principle. *Studies in History and Philosophy of Modern Physics*, *27*, 47–79.

van Campenhout, J., & Cover, T. (1981). Maximum entropy and conditional probability. *IEEE Transactions on Information Theory*, *IT-27*(4), 483–489.

van Fraassen, B. C. (1981). A problem for relative information minimizers. *British Journal for the Philosophy of Science*, *32*, 375–379.

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

vos Savant, M. (May 26, 1996). Ask Marilyn. *Parade Magazine*. There were also followup articles in *Parade Magazine* on Dec. 1, 1996, March 30, 1997, July 27, 1997, and October 19, 1997.

vos Savant, M. (Sept. 9, 1990). Ask Marilyn. *Parade Magazine*, 15. Follow-up articles appeared in *Parade Magazine* on Dec. 2, 1990 (p. 25) and Feb. 17, 1991 (p. 12).