Statistical Science
2021, Vol. 36, No. 2, 169–190
https://doi.org/10.1214/19-STS765

Judicious Judgment Meets Unsettling Updating: Dilation, Sure Loss and Simpson's Paradox[1]

Ruobin Gong and Xiao-Li Meng

Abstract. Imprecise probabilities alleviate the need for high-resolution and unwarranted assumptions in statistical modeling. They present an alternative strategy to reduce irreplicable findings. However, updating imprecise models requires the user to choose among alternative updating rules. Competing rules can result in incompatible inferences, and exhibit *dilation*, *contraction* and *sure loss*, unsettling phenomena that cannot occur with precise probabilities and the regular Bayes rule. We revisit some famous statistical paradoxes and show that the logical fallacy stems from a set of marginally plausible yet jointly incommensurable model assumptions akin to the trio of phenomena above. Discrepancies between the generalized Bayes ($\mathfrak{B}$) rule, Dempster's ($\mathfrak{D}$) rule and the Geometric ($\mathfrak{G}$) rule as competing updating rules for Choquet capacities of order 2 are discussed. We note that (1) $\mathfrak{B}$-rule cannot contract nor induce sure loss, but is the most prone to dilation due to "overfitting" in a certain sense; (2) in absence of prior information, both $\mathfrak{B}$- and $\mathfrak{G}$-rules are incapable to learn from data however informative they may be; (3) $\mathfrak{D}$- and $\mathfrak{G}$-rules can mathematically contradict each other by contracting while the other dilating. These findings highlight the invaluable role of judicious judgment in handling low-resolution information, and the care that needs to be take when applying updating rules to imprecise probability models.

Key words and phrases: Imprecise probability, model uncertainty, Choquet capacity, belief function, coherence, Monty Hall problem.

1. THERE IS NO FREE LUNCH

Statistical learning is a process through which models perform updates in light of new information, according to a prespecified set of operation rules. As new observations arrive, a good statistical model revises and adapts its uncertainty quantification according to what has just been observed. If a model a priori judges the probability of an event A to be $P(A)$, after learning event B happened, it may update the posterior probability according to the Bayes rule:

$$P(A \mid B) = P(A)\frac{P(B \mid A)}{P(B)}.$$

Exactly one of three things will happen: $P(A \mid B) > P(A)$, $P(A \mid B) < P(A)$ or $P(A \mid B) = P(A)$. Moreover, $P(A \mid B) > P(A)$ if and only if $P(A \mid B^c) < P(A)$, that is, if B expresses positive support for A, its complement must express negative support. The comparison of prior and posterior probabilities of A encapsulates its *association* with the observed evidence B, a fundamental characterization of the contribution made by a piece of statistical information.

Nevertheless, there exist modeling situations in which associations do not comply with our well-founded intuition. We sketch a series of such examples, well known from textbook probability problems to real-life statistical inference, which will serve as the basis of our analysis throughout the paper. Many of them, known as paradoxes, bear multiple solutions that have long been the center of dispute and explication in the literature. What makes all of

Ruobin Gong is Assistant Professor of Statistics, Rutgers University, 110 Frelinghuysen Road, Piscataway, New Jersey, 08854, USA (e-mail: rg915@stat.rutgers.edu). Xiao-Li Meng is Whipple V. N. Jones Professor of Statistics, Harvard University, 1 Oxford Street, Cambridge, Massachusetts, 02138, USA (e-mail: meng@stat.harvard.edu).

[1]Discussed in 10.1214/21-STS765A, 10.1214/21-STS765B, 10.1214/21-STS765C, 10.1214/21-STS765D; rejoinder at 10.1214/21-STS765REJ.

them thought provoking is the apparent change from prior to posterior judgments of an event of interest that most will find counterintuitive. That, as we will see, is a consequence of the ambiguity in the probabilistic specification of the model itself, ambiguity that cannot be meaningfully resolved by any automated rule.

1.1 Statistical Paradox or Imprecise Probability?

EXAMPLE 1 (Treatment efficacy before and after randomization; Section 2.2). Patients Oreta and Tang are participating in a clinical trial, in which one of them will receive treatment I, and the other treatment II, with equal probability. Let A denote the event that Oreta will improve more from this trial than Tang (assuming no ties), and let B denote the event that Tang is assigned to treatment I. Before the treatment is assigned, we clearly have $P(A) = 1/2$ because the situation is fully symmetric (in the absence of any other information). However, after the assignment is observed, we seem to have no good idea of the value of either $P(A \mid B)$ or $P(A \mid B^c)$, other than they are both bounded within $[0, 1]$.

Example 1 showcases a severe form of "confusion" expressed by the model as the prior probability updates to posterior probability in light of *any* new information. The precise prior judgments $P(A) = 1/2$ and $P(A^c) = 1/2$ are both bound to suffer a loss of precision by the sheer act of conditioning on any event in $\mathcal{B} = \{B, B^c\}$. A central topic of this paper is the *dilation* phenomenon, revealed by Good (1974) and investigated in depth by Seidenfeld and Wasserman (1993), Herron, Seidenfeld and Wasserman (1994, 1997), Pedersen and Wheeler (2014). A formal definition is given in Section 3.1.

EXAMPLE 2 (The boxer, the wrestler and the coin flip (Gelman, 2006); Sections 3.1 and 6.2). The greatest boxer and the greatest wrestler are scheduled to fight. Who will defeat the other? Let $Y = 1$ if the boxer wins; $Y = 0$ if the wrestler wins. Also, let $X = 1$ if a toss of a fair coin yields heads; $X = 0$ if tails. A witness at both the fighting match and the coin flip tells us that $X = Y$. Given this, what is the boxer's chance of winning, $P(Y = 1 \mid X = Y)$?

EXAMPLE 3 (Three prisoners (Diaconis, 1978, Diaconis and Zabell, 1983); Sections 3.2 and 6.3). Three death row inmates A, B and C are told, on the night before their execution, that one of them has been chosen at random to receive parole, but it will not be announced until the next morning. Desperately hoping to learn immediately, prisoner A says to the guard: "Since at least one of B and C will be executed, you will give away no information about my own chance by giving the name of just one of either B or C who is going to be executed." Convinced of this argument, the guard truthfully says, "B will be executed." Given this information, how should A judge his living prospect, $P(A \text{ lives} \mid \text{guard says } B)$?

EXAMPLE 4 (Simpson's paradox (Simpson, 1951, Blyth, 1972); Section 5). We would like to evaluate the effectiveness of a novel treatment (experimental) compared to its standard counterpart (control). Let $Z = 1$ denote assignment of the experimental treatment, 0 the control treatment, and let $Y = 1$ denote the event of a recovery, 0 otherwise. Let $U \in \{1, 2, \ldots, K\}$ be a covariate of the patients, a K-level categorical indicator variable. One could imagine K to be very large, to the extent that the univariate U creates sufficiently individualized strata among the patient population.

Suppose we learn from reliable clinical studies that the experimental treatment works better than the control for all K subtypes of patients. That is, for $k = 1, \ldots, K$,

$$p_k \equiv P(Y = 1 \mid Z = 1, U = k) > q_k \equiv P(Y = 1 \mid Z = 0, U = k). \tag{1.1}$$

Nevertheless, field studies consisting of feedback reports from clinics and hospitals seem to suggest otherwise; that on an overall basis, the control treatment cures more patients than the experimental treatment. That is,

$$\bar{p}_{\text{obs}} \equiv P_{\text{obs}}(Y = 1 \mid Z = 1) < \bar{q}_{\text{obs}} \equiv P_{\text{obs}}(Y = 1 \mid Z = 0). \tag{1.2}$$

How do we resolve the apparent conflict between the conditional inference in (1.1) and the marginal inference in (1.2)?

The above examples will be examined in detail in Sections 2 through 4. All of them, despite disguised with cunning descriptions, share the characteristic of an *imprecise model*. Their narratives imply the existence of a joint distribution, yet only a subset of marginal information is precisely specified.

For instance, in Example 1, while the treatment assignment (B) is known to be fair prior to randomization, the improvement event A is not measurable with respect to the B margin, effectively posing a Fréchet class of joint distributions on the $\{A, B\}$ space. The only statements we can make about $P(A \mid *)$ are the trivial bounds that $0 \le P(A \mid *) \le 1$, whether $*$ is B or B^c, leading to the dilation phenomenon. In Example 2, the coin margin X is fully known a priori, but the relationship between the fighters Y and the coin X, crucial for quantifying the event $\{X = Y\}$, is unspecified. In Example 3, the guard's tendency to report B over C is unspecified in the case that A was granted parole, yet A's survival probability depends critically on this reporting tendency. In Example 4, the joint specification of $\{Z, U\}$ is missing, and that happens to be key to the seemingly paradoxical reversal effect. In all of these examples, the water is muddied by an unspecified but necessary piece of relational knowledge, which in turn imposes on the modeler a choice among a multiplicity of updating rules, each supplying a distinct set of assumptions to complement this ambiguity.

1.2 What do We Try to Accomplish in This Paper?

Unsettling phenomena to be discussed in this paper reflect unusual ways through which more information can seemingly "harm" our existing knowledge of the state of matters. These phenomena are not foreign to statisticians, but are seen as anomalies or even paradoxes, far from everyday model building. In fact, whenever there is a fully and precisely specified probability model, none of these phenomena would occur. Would not we all be safer then by staying away from any imprecise model? Quite the contrary, we argue. Imprecise models are unavoidable even in basic statistical modeling, and sometimes they are disguised as precise models only to trick us into blindness. Simpson's paradox, re-examined in Section 5, is one of such cases. Without acknowledging the imprecise nature of modeling, one is ill-suited to make judicious choices among the updating rules and treatments of evidence.

We aim to investigate these perceived anomalies as they occur during the updating of imprecise models, and their implications on the choice of updating rules. Imprecise models in statistical modeling are ubiquitous and can be easily induced from precise models through the introduction of external variables. When model imprecision is present, a choice among updating rules is a necessity, and it reflects the modeler's judgment on how statistical evidence at hand should be used. With the recent surge of interest in imprecise probability-based and related statistical frameworks including generalized Fiducial inference (Hannig et al., 2016), confidence distribution (Hannig and Xie, 2012, Xie and Singh, 2013, Schweder and Hjort, 2016) and inferential models (Martin and Liu, 2016), we are compelled to bring attention to the nonnegligible choice of combining and conditioning rules for statistical evidence.

The remainder of this paper starts with an introduction to some formal notation of imprecise probabilities in Section 2.1, particularly of Choquet capacities of order 2 as well as belief functions, a versatile special case which can also be formulated as a precise model for imprecise states, that is, set-valued random variables. Three main updating rules are introduced in Section 2.2, all of which are applicable to Choquet capacities of order 2. Section 3 defines dilation, contraction and sure loss as phenomena that happen during imprecise model updating, and Section 4 compares and contrasts the behavior of the three updating rules, especially as they exhibit dilation and sure loss, and illustrates them with an additional example. Section 5 extends the discussion from conditioning rules to marginalizing rules by showing how Simpson's paradox is a consequence of an ill-chosen updating rule that induces sure loss in aggregation. It also shows how imprecise models can be easily induced from precise ones. When do the updating rules differ, and how? We believe these questions will shed light on the means through which information could contribute to imprecise statistical models, a topic we discuss in Section 6, among others.

2. IMPRECISE PROBABILITIES AND THEIR UPDATING RULES

2.1 Lower and Upper Probabilities

This section introduces formal concepts and notation for imprecise probability needed within the scope of this paper. Readers who are familiar with the notions of lower and upper probabilities, Choquet capacity and belief function may skip to Section 2.2.

DEFINITION 2.1 (Lower and upper probabilities). Let Ω be a separable and completely metrizable space, $\mathscr{B}(\Omega)$ its Borel σ-algebra and $\mathscr{M}$ the set of all probability measures on Ω. The *lower and upper probabilities* of a set of probability measures $\Pi \subset \mathscr{M}$ are set functions

$$\underline{P}(A) = \inf_{P\in\Pi} P(A), \quad \text{and} \quad \overline{P}(A) = \sup_{P\in\Pi} P(A),$$

for all $A \in \mathscr{B}(\Omega)$. $\underline{P}$ and $\overline{P}$ are *conjugate* in the sense that $\overline{P}(A) = 1 - \underline{P}(A^c)$.

The conjugacy of $\underline{P}$ and $\overline{P}$ means that knowing one is sufficient for knowing the other. We may refer to either one individually with the understanding of their one-to-one relationship. Next, we introduce Choquet capacities, an important class of imprecise probabilities widely used in robust statistics (Huber and Strassen, 1973).

DEFINITION 2.2 (Choquet capacities of order k). Suppose $\underline{P}$ is a lower probability such that $\{P \in \mathscr{M}; P \geq \underline{P}\}$, the set of probability measures *compatible* with $\underline{P}$ is relatively compact .[1] $\underline{P}$ is a *Choquet capacity of order* k, or k*-monotone* capacity, if for every Borel-measurable collection of $\{A, A_1, \ldots, A_k\}$ such that $A_i \subset A$ for all $i = 1, \ldots, k$, we have

$$\underline{P}(A) \geq \sum_{\varnothing\neq I\subset\{1,\ldots,k\}} (-1)^{|I|-1} \underline{P}\Big(\bigcap_{i\in I} A_i\Big), \tag{2.1}$$

where $|S|$ denotes the number of elements in the set S. Its conjugate capacity function $\overline{P}$ is a called a k*-alternating* capacity, because it satisfies for every Borel-measurable collection of $\{A, A_1, \ldots, A_k\}$ such that $A \subset A_i$ for all $i = 1, \ldots, k$,

$$\overline{P}(A) \leq \sum_{\varnothing\neq I\subset\{1,\ldots,k\}} (-1)^{|I|-1} \overline{P}\Big(\bigcup_{i\in I} A_i\Big). \tag{2.2}$$

[1] A set of probability measures Π on $(\Omega, \mathscr{B}(\Omega))$ is relative compact if every sequence of elements of Π contains a weakly convergent subsequence. By Prokhorov's theorem, Π is relatively compact if and only if it is tight. See Chapter 1.5 of Billingsley (2013).

If a Choquet capacity is $(k+1)$-monotone, it is k-monotone as well. The smaller the k, the broader the class. In particular, Choquet capacities of order 2 satisfy $\underline{P}(A \cup B) \geq \underline{P}(A) + \underline{P}(B) - \underline{P}(A \cap B)$ for all $A, B \in \mathscr{B}(\Omega)$. A most special case of Choquet capacity is belief function (Shafer, 1979).

DEFINITION 2.3 (Belief function). $\underline{P}$ is called a *belief function* if it is a Choquet capacity of order ∞, that is, if (2.1), and hence (2.2) hold for every k.

Precise probabilities are a special type of belief function. Indeed, one of the probability axioms requires that the inequality (2.1) hold with equality for all countable collections of sets $\{A, A_1, A_2 \cdots\}$ when $A = \bigcup_i A_i$. In turn, belief functions make up only a small class of imprecise probabilities, with their own specializations and limitations when it comes to characterizing uncertain knowledge. Pearl (1990) noted that many imprecise probabilities expressed in conditional forms, a category in which Examples 1 and 4 falls, cannot be fully captured by belief functions. On the other hand, belief functions are versatile in that they possess a second interpretation as a precise probability distribution over the subsets of Ω. In other words, just as a probability function induces a (point-valued) random variable on Ω itself, a belief function induces a *set-valued* random variable on the power set of Ω. This point is made clear in the next definition.

DEFINITION 2.4 (Mass function of a belief function). Suppose Ω is finite, and $\underline{P}$ is a belief function on Ω. The *mass function* associated with $\underline{P}$ is the nonnegative set function $m : \mathscr{P}(\Omega) \to [0, 1]$ such that

$$m(A) = \sum_{B \subseteq A} (-1)^{|A-B|} \underline{P}(B), \quad \text{for all } A \in \mathscr{B}(\Omega) \tag{2.3}$$

where $A - B = A \cap B^c$. The mass function m is uniquely determined by $\underline{P}$, and satisfies (1) $m(\varnothing) = 0$, (2) $\sum_{A \subseteq \Omega} m(A) = 1$, and (3) $\underline{P}(A) = \sum_{B \subseteq A} m(B)$.

Formula (2.3), called the *Möbius transform* of $\underline{P}$ (Yager and Liu, 2008), specifies a precise probability distribution over the subsets of Ω. Definition 2.4 is applicable to finite Ω, suitable for our discussion of Examples 2 and 3 in Section 3 as well as Example 5 in Section 4.5. Definitions for infinite Ω can be obtained upon introducing extra regularity conditions (Nguyen, 1978, Shafer, 1979), which we will not go into in this paper.

2.2 Updating Rules for Lower and Upper Probabilities

To update a set of probabilities Π *given* a set $B \in \mathscr{B}(\Omega)$ is to replace the set function $\underline{P}$ with a version of the conditional set function $\underline{P}_\bullet(\cdot \mid B)$. The definition of $\underline{P}_\bullet$ is given by the updating rule. We emphasize that, to say an event is "given" does not necessarily mean it is observed. In hypothetical contemplations, we often employ conditional statements about all events in a partition, for example, $\mathcal{B} = \{B, B^c\}$, even if logically we cannot observe B and B^c simultaneously. Therefore, the phrase "given" should be understood as imposing a mathematical constraint derived from B. When Π contains a single, precise statistical model, the Bayes rule entirely dictates how we use the information supplied by B. But when Π is imprecise and does not possess *sharp* knowledge about B, that is, $\underline{P}(B) < \overline{P}(B)$ (Dempster, 1967), the updating rule itself becomes an imprecise matter. As a consequence, there exists multiple reasonable ways to use the information B. For example, whether B *supports* an assertion A and whether B *fails to contradict* A are two different criteria for admissible evidence. This raises both flexibility and confusion in defining the updating rules. Here, we supply the formal definitions of three viable updating rules for lower and upper probabilities: the *generalized Bayes* rule, *Dempster's* rule and the *Geometric* rule. Important differences and relationships exist among these rules, as we shall present in Section 4.

To define the generalized Bayes rule, we recall Example 1. Using the notation in 2.1, we rewrite the imprecise model in terms of its prior upper and lower probabilities of event A, which are precisely one half: $\underline{P}(A) = \overline{P}(A) = 0.5$. The question is: what are the upper and lower probabilities of A *given* the treatment assignments in $\mathcal{B} = \{B, B^c\}$? For example, the answer could be

$$\underline{P}_{\mathfrak{B}}(A \mid B) = 0, \quad \overline{P}_{\mathfrak{B}}(A \mid B) = 1 \quad \text{and}$$
$$\underline{P}_{\mathfrak{B}}(A \mid B^c) = 0, \quad \overline{P}_{\mathfrak{B}}(A \mid B^c) = 1.$$

The expressions $\underline{P}_{\mathfrak{B}}$ and $\overline{P}_{\mathfrak{B}}$, where the subscript $\mathfrak{B}$ is for $\mathfrak{B}$*ayes*, signify the use of the generalized Bayes rule, as defined below.

DEFINITION 2.5 (Generalized Bayes rule). Let Π be a convex and closed set of probability measures on Ω (with respect to the total variation topology, as in Seidenfeld and Wasserman, 1993). The conditional lower and upper probabilities according to the *generalized* $\mathfrak{B}$*ayes rule* are set functions $\underline{P}_{\mathfrak{B}}$ and $\overline{P}_{\mathfrak{B}}$ such that, for $A, B \in \mathscr{B}(\Omega)$,

$$\underline{P}_{\mathfrak{B}}(A \mid B) = \inf_{P \in \Pi} \frac{P(A \cap B)}{P(B)}, \tag{2.4}$$

$$\overline{P}_{\mathfrak{B}}(A \mid B) = \sup_{P \in \Pi} \frac{P(A \cap B)}{P(B)}. \tag{2.5}$$

That is, the conditional lower and upper probabilities are respectively the minimal and maximal Bayesian conditional probability among elements of Π. In their definition, Seidenfeld and Wasserman (1993) required that $\underline{P}(B) > 0$, which guarantees $P(B) > 0$ for all $P \in \Pi$.

This guarantees that the ratios in (2.4) and (2.5) are always well defined.

The generalized Bayes rule is a most widely employed updating rule for coherent lower and upper probabilities (Walley, 1991), and notable for exhibiting dilation. In Example 1, as a consequence of employing the rule, the conclusion appears puzzling: Tang will surely receive one of the two treatments, and one would expect that, in the worst case scenario, learning about the treatment assignment is completely useless, that is, having no effect on our a priori assessment of $P(A)$. But how could it be that the knowledge of something can do more harm than being useless?

To better understand the behavior of the generalized Bayes rule, we now present two alternative updating rules for sets of probabilities as means of comparison. Both Dempster's rule of conditioning and the Geometric rule were originally proposed for use with the special case of belief functions; however, their expressions compose intriguing counterparts to the generalized Bayes rule. Section 4 is dedicated to a comparison among the trio of rules.

Dempster's rule of conditioning is central to the Dempster–Shafer theory of belief functions (Dempster, 1967, Shafer, 1976). The conditioning operation is a special case of Dempster's rule of combination, equivalent to combining one belief function with another that puts 100% mass on one particular subset, $B \in \mathscr{B}(\Omega)$, on which we wish to condition. Specifically, let $\underline{P}$ be a belief function such that $\underline{P}(B) > 0$, and m be its associated mass function given by (2.3). Let $\underline{P}_0$ be a separate belief function such that its associated mass function $m_0(B) = 1$. The conditional belief function $\underline{P}_{\mathfrak{D}}(\cdot \mid B)$ is defined as

$$\underline{P}_{\mathfrak{D}}(A \mid B) = \underline{P}(A) \oplus \underline{P}_0(B), \quad \text{for all } A \in \mathscr{B}(\Omega),$$

where the combination operator "$\oplus$" is defined in Shafer (1976) to imply that the mass function associated with $\underline{P}_{\mathfrak{D}}(\cdot \mid B)$ is

$$m_{\mathfrak{D}}(A \mid B) = \frac{\sum_{C \cap B = A} m(C)}{\sum_{C' \cap B \neq \varnothing} m(C')}, \quad \text{for all } A \in \mathscr{B}(\Omega). \tag{2.6}$$

Consequently, Dempster's rule of conditioning yields the following form.

DEFINITION 2.6 (Dempster's rule of conditioning). Let $\underline{P}$ be a belief function over Ω, and Π the set of probabilities compatible with $\underline{P}$ (in the sense of Definition 2.2). The lower and upper probabilities according to $\mathfrak{D}$*empster's rule of conditioning* are set functions $\underline{P}_{\mathfrak{D}}$ and $\overline{P}_{\mathfrak{D}}$ such that for $A, B \in \mathscr{B}(\Omega)$ with $\overline{P}(B) > 0$,

$$\underline{P}_{\mathfrak{D}}(A \mid B) = 1 - \overline{P}_{\mathfrak{D}}(A^c \mid B), \tag{2.7}$$

$$\overline{P}_{\mathfrak{D}}(A \mid B) = \frac{\sup_{P \in \Pi} P(A \cap B)}{\sup_{P \in \Pi} P(B)}. \tag{2.8}$$

Hence $\overline{P}_{\mathfrak{D}}(A \mid B)$ differs from $\overline{P}_{\mathfrak{B}}(A \mid B)$ of (2.5) by taking the ratio of the suprema, instead of the supremum of the ratio $P(A \cap B)/P(B)$. An operational view of (2.8) is helpful for understanding exactly what information is retained by Dempster's rule (Gong and Meng, 2021). Denote by $\mathcal{R}$ the set-valued random variable whose distribution is dictated by the mass function corresponding to $\underline{P}$. Dempster's rule of conditioning of $\underline{P}$ on set B is akin to applying a B-shaped "cookie cutter" to all realizations of $\mathcal{R}$. It retains all the nonempty intersections $B \cap \mathcal{R}$, and defines the associated conditional mass function $m_{\mathfrak{D}}(\cdot \mid B)$ according to (2.6), that is, renormalizing m among the $\mathcal{R}$'s pertinent to the retained sets. The functional form of (2.8) reveals that, Dempster's upper conditional probability admits evidence to its numerator and denominator, both according to whether the evidence *fails to contradict* $A \cap B$ and B. This stands in contrast to the Geometric rule proposed by Suppes and Zanotti (1977), as defined below.

DEFINITION 2.7 (The Geometric rule). Let $\underline{P}$ be a belief function as in Definition 2.6. The conditional lower and upper probabilities according to the $\mathfrak{G}$*eometric rule* are set functions $\underline{P}_{\mathfrak{G}}$ and $\overline{P}_{\mathfrak{G}}$ such that for $A, B \in \mathscr{B}(\Omega)$ with $\underline{P}(B) > 0$,

$$\underline{P}_{\mathfrak{G}}(A \mid B) = \frac{\inf_{P \in \Pi} P(A \cap B)}{\inf_{P \in \Pi} P(B)}, \tag{2.9}$$

$$\overline{P}_{\mathfrak{G}}(A \mid B) = 1 - \underline{P}_{\mathfrak{G}}(A^c \mid B). \tag{2.10}$$

Mathematically, the Geometric rule is a dual to Dempster's rule by replacing the latter's suprema for upper probability in (2.8) with the infima for lower probability in (2.9). Viewed as a set operation, the Geometric rule differs from Dempster's rule in that it only retains $\mathcal{R}$ if fully contained within B, and renormalizes the mass function among the retained sets. Looking at (2.9), the Geometric lower conditional probability admits evidence to its numerator and denominator, both according to whether the evidence *supports* $A \cap B$ or B.Section 4 further describes some relationships between the two rules.

Just like the generalized Bayes rule, both Dempster's and the Geometric rules suffer from updating anomalies. In his review of Shafer (1976), Diaconis (1978) discussed a paradoxical conclusion for the three prisoners example (reproduced here as Example 3) using Dempster's rule, and inquired about the option of the Geometric rule as an alternative rule of updating. As we will show in Section 3.2, the Geometric rule does no better job than Dempster's rule for this paradox, as in fact both rules exhibit the *sure loss* phenomenon. More updating rules for belief functions exist beyond Dempster's and the Geometric rule, including the disjunctive rule by Smets (1993) based on set union operations, the open-world conjunctive rule which is the unnormalized version of Dempster's rule as employed in the transferable belief models, as well as

others, for example, Yager (1987), Kohlas (1991), Kruse and Schwecke (1990). Smets (1991) provided a broad overview of an array of updating rules.

2.3 IP Updating Rules Are Not Pure Conditional Probabilities

A key distinction between the updating rules for imprecise probabilities and the Bayes rule for precise probabilities is that the former does not follow pure conditional probability calculations, but rather a mixture of probability and bound-seeking operations. This is most easily seen in the following expressions obtained by Fagin and Halpern (1991) for generalized Bayes rule:

$$\underline{P}_{\mathfrak{B}}(A \mid B) = \frac{\underline{P}(A \cap B)}{\underline{P}(A \cap B) + \overline{P}(A^c \cap B)}, \tag{2.11}$$

$$\overline{P}_{\mathfrak{B}}(A \mid B) = \frac{\overline{P}(A \cap B)}{\overline{P}(A \cap B) + \underline{P}(A^c \cap B)}. \tag{2.12}$$

Compared to the familiar Bayes formula

$$\begin{aligned} P(A \mid B) &= \frac{P(A \cap B)}{P(A \cap B) + P(A^c \cap B)} \\ &\equiv \frac{P(A \cap B)}{P(B)}, \end{aligned} \tag{2.13}$$

we see that the generalized Bayes rule not only replaces P by $\overline{P}$ or $\underline{P}$, but it also mixes them in one expression. This mean that in general, the "conditional probability" obtained by the generalized Bayes rule is not a genuine probability under a single probability distribution. Worse, the distributions which attain the extrema, $\overline{P}(S)$ or $\underline{P}(S)$, in general depends on S itself. This is a clear case of "overfitting," as probabilities are "cherry-picked" to make S most or least likely.

One might attempt to fix the mixing issue by replacing the right-hand sides in (2.11) and (2.12), respectively, by

$$\frac{\underline{P}(A \cap B)}{\underline{P}(A \cap B) + \underline{P}(A^c \cap B)} \quad \text{and} \quad \frac{\overline{P}(A \cap B)}{\overline{P}(A \cap B) + \overline{P}(A^c \cap B)}. \tag{2.14}$$

However, $\underline{P}(A \cap B) + \underline{P}(A^c \cap B)$ generally is smaller than $\underline{P}(B)$ because one may be sure that a state is in B, but unsure if it is in $A \cap B$ or $A^c \cap B$. Indeed, $\underline{P}(A \cap B) + \underline{P}(A^c \cap B)$ can be zero, while $\underline{P}(B) > 0$. Similarly, we can have $\overline{P}(A \cap B) + \overline{P}(A^c \cap B) > 1$ because evidence that does not contradict $A \cap B$ or $A^c \cap B$ get double counted in the sum $\overline{P}(A \cap B) + \overline{P}(A^c \cap B)$.

These observations should remind us that while the expressions in (2.14) may appear to be natural generalizations of the Bayes formula in the middle expression of (2.13), they are not legit probabilistic quantities even in the context of imprecise probability (e.g., no imprecise probability can exceed one). Consequently, it makes more sense to directly use $\underline{P}(B)$ or $\overline{P}(B)$ to replace $P(B)$ in the right-hand expression of (2.13). The results are exactly the Geometric rule:

$$\underline{P}_{\mathfrak{G}}(A \mid B) = \frac{\underline{P}(A \cap B)}{\underline{P}(B)}, \tag{2.15}$$

and the Dempster's rule:

$$\overline{P}_{\mathfrak{D}}(A \mid B) = \frac{\overline{P}(A \cap B)}{\overline{P}(B)}. \tag{2.16}$$

Expression (2.15) makes it clear that the Geometric rule endorses a stringent interpretation of what counts as evidence for both the query (A) and conditioning (B) events, by admitting only evidence that *supports* its constituents into the lower conditional probability. Similarly, (2.16) shows that Dempster's rule endorses a lenient interpretation of both parts, by permitting all evidence that *does not contradict* into the upper conditional probability.

In contrast, generalized Bayes rule optimizes not over the space of admissible evidence, but over the set of all conditional probabilities implied by the prior imprecise model. The expressions (2.11) and (2.12) reveal that, compared to (2.16) and (2.15), the implied criteria of what counts as admissible evidence is disparate for the query and conditioning events on the numerator versus the denominator. This results in the aforementioned "overfitting" phenomenon, a point to which we will return in Section 3.2.

2.4 Generalizations to Choquet Capacities

The generalized Bayes rule was designed to work with sets of convex and closed probabilities, of which those sets of probabilities generated by Choquet capacities of order 2 are a special case. It has been shown that, when applied to prior sets of probabilities that are Choquet capacities of order 2, the posterior sets of probabilities by the generalized Bayes rule remain in the class (Walley, 1981, Wasserman and Kadane, 1990). That is, Choquet capacities of order 2 are closed with respect to the generalized Bayes rule. A natural question then if this property holds for Dempster's rule or the Geometric rule. The next theorem shows that the answer is yes: Choquet capacities of order k, for any $k \geq 2$, are closed with respect to both rules.

THEOREM 2.1. *Let $\underline{P}$ be a k-monotone Choquet capacity on Ω, and event B such that the set functions $\underline{P}_{\mathfrak{D}}(\cdot \mid B)$ in (2.7) and $\underline{P}_{\mathfrak{G}}(\cdot \mid B)$ in (2.9) are well defined. Then $\underline{P}_{\mathfrak{D}}(\cdot \mid B)$ and $\underline{P}_{\mathfrak{G}}(\cdot \mid B)$ are both k-monotone.*

PROOF. To say $\underline{P}$ is k-monotone implies for all Borel-measurable collections $\{A_1, \ldots, A_k\}$,

$$\begin{aligned} \underline{P}\left(\bigcup_{i=1}^{k} A_i\right) \geq \sum_{i=1}^{k} \underline{P}(A_i) - \sum_{i<j} \underline{P}(A_i \cap A_j) \\ + \cdots + (-1)^{k+1} \underline{P}\left(\bigcap_{i=1}^{k} A_i\right) \end{aligned}$$

or, equivalently, $\overline{P}$ is k-alternating:

$$\overline{P}\left(\bigcap_{i=1}^{k} A_i\right) \le \sum_{i=1}^{k} \overline{P}(A_i) - \sum_{i<j} \overline{P}(A_i \cup A_j) + \cdots + (-1)^{k+1} \overline{P}\left(\bigcup_{i=1}^{k} A_i\right).$$

For Dempster's rule, we have

$$\begin{aligned}
&\overline{P}_{\mathfrak{D}}\left(\bigcap_{i=1}^{k} A_i \mid B\right) \\
&\quad = \frac{\overline{P}((\bigcap_{i=1}^{k} A_i) \cap B)}{\overline{P}(B)} = \frac{\overline{P}(\bigcap_{i=1}^{k} (A_i \cap B))}{\overline{P}(B)} \\
&\quad \le \frac{1}{\overline{P}(B)} \cdot \left[\sum_{i=1}^{k} \overline{P}(A_i \cap B) \right. \\
&\qquad - \sum_{i<j} \overline{P}\big((A_i \cap B) \cup (A_j \cap B)\big) + \cdots \\
&\qquad \left. + (-1)^{k+1} \overline{P}\left(\bigcup_{i=1}^{k} (A_i \cap B)\right)\right] \\
&\quad = \sum_{i=1}^{k} \overline{P}_{\mathfrak{D}}(A_i \mid B) - \sum_{i<j} \overline{P}_{\mathfrak{D}}(A_i \cup A_j \mid B) + \cdots \\
&\qquad + (-1)^{k+1} \overline{P}_{\mathfrak{D}}\left(\bigcup_{i=1}^{k} A_i \mid B\right).
\end{aligned}$$

Similarly, for the Geometric rule,

$$\begin{aligned}
&\underline{P}_{\mathfrak{G}}\left(\bigcup_{i=1}^{k} A_i \mid B\right) \\
&\quad = \frac{\underline{P}((\bigcup_{i=1}^{k} A_i) \cap B)}{\underline{P}(B)} = \frac{\underline{P}(\bigcup_{i=1}^{k} (A_i \cap B))}{\underline{P}(B)} \\
&\quad \ge \frac{1}{\underline{P}(B)} \cdot \left[\sum_{i=1}^{k} \underline{P}(A_i \cap B) \right. \\
&\qquad - \sum_{i<j} \underline{P}(A_i \cap A_j \cap B) + \cdots \\
&\qquad \left. + (-1)^{k+1} \underline{P}\left(\bigcap_{i=1}^{k} A_i \cap B\right)\right] \\
&\quad = \sum_{i=1}^{k} \underline{P}_{\mathfrak{G}}(A_i \mid B) \\
&\qquad - \sum_{i<j} \underline{P}_{\mathfrak{G}}(A_i \cap A_j \mid B) + \cdots \\
&\qquad + (-1)^{k+1} \underline{P}_{\mathfrak{G}}\left(\bigcap_{i=1}^{k} A_i \mid B\right).
\end{aligned}$$

Hence k-monotonicity is preserved by both Dempster's and the Geometric rules of updating when applied to k-monotone Choquet capacities. □

3. THE UNSETTLING UPDATES IN IMPRECISE PROBABILITIES

An imprecise model permits, and indeed requires, a choice of updating rule. Different choices may exhibit updates with seemingly troubling interpretations, notably *dilation*, *contraction* and *sure loss*. This section supplies an in-depth look at these phenomena. The subscript "•" used in the definitions below is crucial because, given the same imprecise model specification, a phenomenon can be induced by one rule but not by another. The choice among updating rules is inseparable from the choice of assumption of a missing information mechanism, and it would be wrong to think that an observable event, as a mathematical constraint, is taken literatim in imprecise probability conditioning. The operational interpretations of Dempster's rule and the Geometric rule presented in the previous section highlight clearly the different uses, by different rules, of the information in the same event being conditioned upon.

3.1 Dilation and Contraction

DEFINITION 3.1 (Dilation). Let $A \in \mathscr{B}(\Omega)$ and $\mathcal{B}$ be a Borel measurable partition of Ω. Let Π be a convex and closed set of probability measures on Ω, $\underline{P}$ its lower probability function, and $\underline{P}_\bullet$ the conditional lower probability function supplied by the updating rule "•". We say that $\mathcal{B}$ *strictly dilates* A under the •-rule if

$$\begin{aligned}
\sup_{B \in \mathcal{B}} \underline{P}_\bullet(A \mid B) &< \underline{P}(A) \le \overline{P}(A) \\
&< \inf_{B \in \mathcal{B}} \overline{P}_\bullet(A \mid B).
\end{aligned} \tag{3.1}$$

If either (but not both) outer inequality is allowed to hold with equality, we simply say $\mathcal{B}$ *dilates* A under the said updating rule.

Dilation means that the conditional upper and lower probability interval of an event A contains that of the unconditional interval, regardless of which B in the space of possibilities $\mathcal{B}$ is observed. Inference for A, as expressed by the imprecise probabilities under the chosen updating rule, will become strictly less precise regardless of what has been learned. This is commonly perceived as unsettling, because one would expect that learning, at least in *some* situations, ought to help the model deliver sharper inference, reflected in a tighter probability interval. But when dilation happens, it seems that as we learn, knowledge does not accumulate and quite the contrary, diminishes surely.

If dilation is something one finds unsettling, the opposing notion, *contraction*, should be nothing less. Contraction happens when the posterior upper and lower probability interval becomes strictly contained within that of the prior, regardless of what is being learned. If a tighter probability interval symbolizes more knowledge, when contraction happens, it is as if some knowledge is created out of thin air. How could it be that whatever is learned, we could always eliminate a fixed set of values of probability that were a priori considered possible? If we could have eliminated them by a pure thought experiment that can never fail, why would we not have eliminated them a priori? Formally, contraction is defined as follows.

DEFINITION 3.2 (Contraction). Let A, $\mathcal{B}$ and $\underline{P}_\bullet$ be the same as in Definition 3.1. We say that $\mathcal{B}$ *strictly contracts* A under the $\bullet$-rule if

$$\begin{aligned}\underline{P}(A) &< \inf_{B\in\mathcal{B}} \underline{P}_\bullet(A \mid B)\\ &\le \sup_{B\in\mathcal{B}} \overline{P}_\bullet(A \mid B) < \overline{P}(A).\end{aligned} \tag{3.2}$$

If either (but not both) outer inequality is allowed to hold with equality, we simply say $\mathcal{B}$ *contracts* A under the said updating rule.

We now illustrate these two unsettling updating phenomena using Example 2, although we defer the discussion of their interpretations to Section 6.

EXAMPLE 2 CONT. (The boxer, the wrestler and the coin flip). By the setup of the model, we know precisely that the coin is fair:

$$P(X=0) = P(X=1) = 1/2. \tag{3.3}$$

However, no information is available about either fighter's chance of winning. That is, if we assume the probability of a boxer's win $P(Y=1) = p_1$, p_1 is allowed to vary between $[0, 1]$. Then according to the imprecise model,

$$\underline{P}(Y=1) = 0, \quad \overline{P}(Y=1) = 1 \tag{3.4}$$

and similarly so for the wrestler's win: $\underline{P}(Y=0) = 0$, $\overline{P}(Y=0) = 1$. The known probabilistic margins specify a belief function, as displayed in Table 1.

When told $X = Y$, how should the model at hand be revised? Two aspects are worth noting:

TABLE 1
Example 2 (boxer and wrestler): mass function representation of the belief function model

	Coin lands heads, either fighter wins $(X,Y) \in \{1\}\times\{0,1\}$	Coin lands tails, either fighter wins $(X,Y) \in \{0\}\times\{0,1\}$
$m(\cdot)$	0.5	0.5

(i) *Posterior inference for the fighters*. As Gelman (2006) noted, Dempster's rule contracts the boxer's chance of winning, because

$$\begin{aligned}\underline{P}_{\mathfrak{D}}(Y=1 \mid X=Y) &= 1/2,\\ \overline{P}_{\mathfrak{D}}(Y=1 \mid X=Y) &= 1/2,\\ \underline{P}_{\mathfrak{D}}(Y=1 \mid X\neq Y) &= 1/2,\\ \overline{P}_{\mathfrak{D}}(Y=1 \mid X\neq Y) &= 1/2,\end{aligned}$$

which are strictly contained within the vacuous prior probability interval as in (3.4). The calculations given the two alternative conditions $X = Y$ and $X \neq Y$ are identical due to symmetry of the setup. In contrast, the generalized Bayes rule cannot contract vacuous prior interval, in this example (see below) and in general (see Theorem 4.8).

(ii) *Posterior inference for the coin*. The generalized Bayes rule dilates the precise a priori information (3.3) on the coin's chance of coming up heads, because

$$\begin{aligned}\underline{P}_{\mathfrak{B}}(X=1 \mid X=Y) &= 0,\\ \overline{P}_{\mathfrak{B}}(X=1 \mid X=Y) &= 1,\\ \underline{P}_{\mathfrak{B}}(X=1 \mid X\neq Y) &= 0,\\ \overline{P}_{\mathfrak{B}}(X=1 \mid X\neq Y) &= 1.\end{aligned}$$

In contrast, Dempster's intervals remain identical to that of the prior interval under either $X = Y$ or $X \neq Y$. Notice that in this example, $\underline{P}(X=Y) = \underline{P}(X\neq Y) = 0$, hence the Geometric rule is not applicable. The generalized Bayes rule in the sense of Seidenfeld and Wasserman (1993) (see Definition 2.5) is not applicable either, however, since the the model is a belief function, we use the result from Fagin and Halpern (1991) as given in (2.11) and (2.12) to obtain the above expressions. This is equivalent to minimizing and maximizing over the restricted sets of probabilities $\{P : P \ge \underline{P}, P(X=Y) > 0\}$ and $\{P : P \ge \underline{P}, P(X\neq Y) > 0\}$, respectively, thus avoiding ill-defined probability ratios.

3.2 Sure Loss

The next type of updating anomaly is even more unsettling, as it is usually regarded as an infringement on the logical coherence of probabilistic reasoning.

DEFINITION 3.3 (Sure loss). Let A, $\mathcal{B}$, $\underline{P}$ and $\underline{P}_\bullet$ be the same as in Definition 3.1. We say that $\mathcal{B}$ incurs *sure loss* in A under the $\bullet$-rule if either

$$\inf_{B\in\mathcal{B}} \underline{P}_\bullet(A \mid B) > \overline{P}(A) \tag{3.5}$$

or

$$\sup_{B\in\mathcal{B}} \overline{P}_\bullet(A \mid B) < \underline{P}(A). \tag{3.6}$$

Sure loss describes a universal and unidirectional displacement of probability judgment before and after conditioning on any event from a subalgebra. That is, after learning anything, the event in question becomes altogether more (or less) likely than before.

The terminology "sure loss" stems from the Bayesian decision-theoretic context, where probabilities are seen to profess personal preferences contingent on which one is willing to make bets. If $\mathcal{B}$ incurs sure loss in A, the beholder of $\underline{P}$ and $\underline{P}_\bullet$ as her personal prior and posterior imprecise probabilities, respectively, can be made to commit a compound bet with a guaranteed negative payoff. To see this, let s, t be two numbers such that

$$\inf_{B\in\mathcal{B}} \underline{P}_\bullet(A \mid B) > s > t > \overline{P}(A).$$

We generate sure loss in the form of (3.5). Since $t > \overline{P}(A)$, I shall accept a bet for which I pay $1-t$, get 1 back if A did not occur and nothing if it did. My expected payoff is $P(A^c) - (1-t) = t - P(A) \geq t - \overline{P}(A) > 0$. On the other hand, since $\underline{P}_\bullet(A \mid B) > s$ for all B, contingent on any B, I shall also accept bets for which I pay s, get 1 back if A did occur and nothing if it did not. Regardless of which B occurs, my expected payoff $P(A \mid B) - s \geq \inf_{B\in\mathcal{B}} \underline{P}_\bullet(A \mid B) - s > 0$. It therefore seems perfectly logical for me to take both bets, as both are expected to have positive return. However, if I do take both bets, then the compound bet is the one with guaranteed payoff of only 1, less than what I have paid for $1 - t + s$ because $s > t$. Therefore, endorsing $\underline{P}_\bullet$ as the updating rule means I am willing to accept a finite collection of bets and certain to lose money, a typical form of incoherent behavior.

Note that if $\mathcal{B}$ incurs sure loss in A in the form of (3.5), it also incurs sure loss in A^c in the form of (3.6), though perhaps the term *sure gain* would be more appropriate—in Émile Borel's words, the former the "imbecile" and the latter the "thief." Whenever a distinction is necessary, we will use the term sure gain in addition to sure loss to highlight the directionality of displacements of posterior probability intervals compared to that of the prior, and will otherwise follow the pessimistic convention (which seems to be a hallmark of statistical or probabilistic terms, such as "risk," "regret," "regression") of the literature and use "sure loss" to refer to both situations if nonambiguous.

We emphasize again that both dilation and sure loss, as concepts describing the change from prior to posterior sets of probabilities, are contingent upon the updating rule. Even with the same imprecise probability model $\underline{P}$, the same partition $\mathcal{B}$ and the same event A, it can well be the case that $\mathcal{B}$ dilates A under one rule and induces sure loss in A under the other. Example 3 below is a situation in which all three rules behave very differently, and Section 4 is dedicated to a characterization of their differential behavior.

TABLE 2
Example 3 (three prisoners): mass function representation of the belief function model

	A lives, guard says {B, C}	B lives, guard says C	C lives, guard says B
$m(\cdot)$	1/3	1/3	1/3

We are now ready to take a careful look at the three prisoners paradox.

EXAMPLE 3 CONT. (Three prisoners). What do we have about the probabilistic model behind the three prisoners? Since exactly one of the three prisoners will receive parole randomly, the prior probabilities of living for each of them are all exact:

$$P(A \text{ lives}) = P(B \text{ lives}) = P(C \text{ lives}) = 1/3.$$

Furthermore, since the guard cannot lie, he has no choice on who to report if the inquirer A does not receive parole. That is,

$$P(\text{guard says } C \mid B \text{ lives}) = P(\text{guard says } B \mid C \text{ lives}) = 1.$$

The above probability specification can be expressed as a belief function model, with mass distribution dictated by the known model margins as represented in Table 2.

We see from the specification that what remains unknown is, in case A indeed receives parole, the propensity of the guard reporting either B or C as dead had he the freedom to choose:

$$\delta_B = P(\text{guard says } B \mid A \text{ lives}) \in [0, 1]. \tag{3.7}$$

As a consequence, the posterior probability of A living is

$$P(A \text{ lives} \mid \text{guard says } B) = \frac{\delta_B}{1+\delta_B}. \tag{3.8}$$

This extra degree of freedom δ_B fully characterizes the set of probabilities implied by the model.

There is a long literature documenting the variety of modes of reasoning to this problem. For example, Mosteller (1965) and Morgan et al. (1991) invoked a similar construction as the δ_B above, in explicating the reasons why many of them are seemingly intuitive yet riddled with logical fallacies. Four types of "popular" answers are reproduced below, reflecting different ways of treating the unknown value δ_B. What's interesting is that, as we will see, three of these answers correspond to those given by the three conditioning rules respectively.

(i) *The indifferentist: assumption of ignorability*. One of the most commonly made assumptions is that the guard has no preference one way or the other about who to report when given the freedom, that is, $\delta_B = 1/2$, thus

$$P(A \text{ lives} \mid \text{guard says } B, \delta_B = 1/2) = 1/3.$$

That is to say, prisoner A would not have benefitted from the knowledge that B is going to be executed, precisely as he claimed to the guard to begin with. The assumption of guard's indifference is equivalent to the *ignorability* assumption commonly employed in the treatment of missing and coarse data (Rubin, 1976, Heitjan and Rubin, 1991, Heitjan, 1994). Despite being intuitive, the assumption is not backed by the model description per se. Neither the posited imprecise model nor the data as reported by the guard can supply any logical evidence to support the ignorability assumption. Therefore, the assertion that ignorability is "intuitive" is a judgment that can be as unreasonable as any other seemingly less intuitive ones, such as the ones below.

(ii) *The optimist: Dempster's rule*. Applying Dempster's rule, we have

$$\underline{P}_{\mathfrak{D}}(A \text{ lives} \mid \text{guard says } B) = 1/2,$$
$$\overline{P}_{\mathfrak{D}}(A \text{ lives} \mid \text{guard says } B) = 1/2.$$

Thus prisoner A felt happier now that his chance of survival increased from $1/3$ to $1/2$. This happiness is gained from assuming the optimistic scenario of $\delta_B = 1$, that is, the guard chose a reporting mechanism that has the highest likelihood given A lives. However, one realizes that the guard could have only reported either B or C, both fully symmetrical in the prior. Had the guard said C would be executed, A would again apply Dempster's rule, thus grow happier following the same logic by effectively assuming $\delta_C = P(\text{guard says } C \mid A \text{ lives}) = 1$. Under the assumption that the guard cannot lie and cannot refuse to answer, $\delta_B + \delta_C = 1$, thus δ_B and δ_C cannot be 1 simultaneously. Hence the reasoning that whatever the guard says, the probability of A living will go up from $1/3$ to $1/2$, which is equivalent to assuming the impossible $\delta_B = \delta_C = 1$, is a direct consequence of a logical fallacy.

(iii) *The pessimist: the Geometric rule*. Applying the Geometric rule, we have

$$\underline{P}_{\mathfrak{G}}(A \text{ lives} \mid \text{guard says } B) = \overline{P}_{\mathfrak{G}}(A \text{ lives} \mid \text{guard says } B) = 0$$

and, by symmetry,

$$\underline{P}_{\mathfrak{G}}(A \text{ lives} \mid \text{guard says } C) = \overline{P}_{\mathfrak{G}}(A \text{ lives} \mid \text{guard says } C) = 0.$$

This answer is perhaps the most striking among all, directly pointing at the absurdity of the assumptions behind the updating rule within this context. Upon hearing anything, prisoner A will deny himself of any hope of living, effectively assuming $\delta_B = 0$ if guard says B and $\delta_C = 0$ if guard says C, two assumptions that are incommensurable with each other because $\delta_B + \delta_C = 1$, much in the same way as the previous case with Dempster's rule.

(iv) *The conservatist: generalized Bayes rule*. The solution suggested by Diaconis (1978), and indeed supplied by the generalized Bayes rule, is

$$\begin{aligned} &\underline{P}_{\mathfrak{B}}(A \text{ lives} \mid \text{guard says } B) = 0, \\ &\overline{P}_{\mathfrak{B}}(A \text{ lives} \mid \text{guard says } B) = 1/2. \end{aligned} \tag{3.9}$$

This answer is a direct consequence of (3.8). As δ_B varies within $[0, 1]$ without any further assumption, one is bound to concur with (3.9). The caveat to it, however, is that again due to prior symmetry of B and C, the generalized Bayes rule will also yield

$$\underline{P}_{\mathfrak{B}}(A \text{ lives} \mid \text{guard says } C) = 0,$$
$$\overline{P}_{\mathfrak{B}}(A \text{ lives} \mid \text{guard says } C) = 1/2.$$

Hence, the generalized Bayes rule results in posterior probability intervals strictly containing the prior probability in all situations.

Our use of the vocabulary "optimism," "pessimism" and "conservatism" to refer to the three updating rules is informed by the interpretation of their respective posterior inference under the effective assumptions they each impose, and is reminiscent of that of Fygenson (2008) for modeling of extrapolated probabilities. These ideological differences illuminate the dynamics among the updating rules for imprecise probability, and highlight the pedagogical significance of the three prisoners' paradox itself. In this example, Dempster's rule updates its conditional lower probability to be greater than that of its prior upper probability thus incurs sure loss of the form (3.5), the Geometric rule behaves the opposite way and incurs sure loss of the form (3.6), and the generalized Bayesian rule exhibits dilation. As far as unsettling updating goes, there seems to be no escape regardless of which rule to choose. How on earth then do we draw a conclusion?

Reading through the literature, the dilated answer supplied by the generalized Bayes rule is the most accepted solution to the paradox. As counterintuitive as it may be, dilation is a professed consequence of an overfitting nature of the generalized Bayes rule, for the rule is inclusive of all possibilities allowed within the ambiguous model, to the point of simultaneously admitting assumptions that are *incommensurable* with one another. As we saw previously, the upper conditional probability $\overline{P}_{\mathfrak{B}}(A \text{ lives} \mid \text{guard says } *) = 1/2$ is achieved under the assumption $\delta_* = 1$, where $*$ can be B or C. Similarly, the lower conditional probability $\underline{P}_{\mathfrak{B}}(A \text{ lives} \mid \text{guard says } *) = 0$ is achieved when $\delta_* = 0$. Since $\delta_C + \delta_B = 1$, δ_C and δ_B cannot simultaneously be 0 or 1. Indeed, when one is 1 the other must be 0. Hence the permissible value of the *pair*

$$\{x = P(A \text{ lives} \mid \text{guard says } B), \quad y = P(A \text{ lives} \mid \text{guard says } C)\}$$

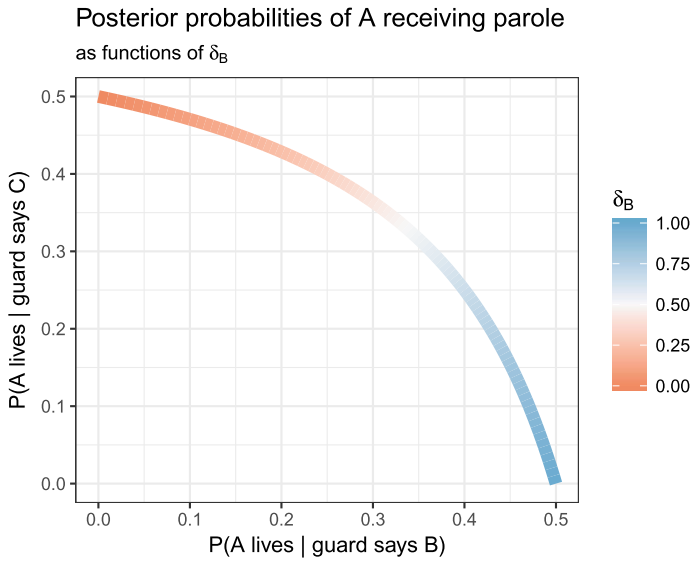

FIG. 1. *Posterior probabilities of prisoner A receiving parole given the guard's two possible answers, as a function of the guard's reporting bias* δ_B (3.7).

forms a one-dimensional curve $y = \frac{1-2x}{2-3x}$ inside the square $[0, 1/2] \times [0, 1/2]$, as depicted in Figure 1. For a given conditioning event $*$, the generalized Bayes rule achieves its extremes by seeking a distribution that itself depends on $*$, namely, a *condition-dependent* conditional distribution $P^{(*)}(\cdot \mid *)$, a clear case of overfitting. Understanding the hidden *incommensurability* is important for preventing logical fallacies such as reasoning under the (wrong) assumption that $\{x, y\}$ can take any value inside the square $[0, 1/2] \times [0, 1/2]$. We will return to the three prisoners again in Section 6.3 to discuss its inferential implications. In particular, the three prisoners' paradox is a direct variant of the Monty Hall problem, which possesses a clean, indisputable decision recommendation.

3.3 What's so Unsettling About Updating Paradoxes?

In case some readers are not yet completely put off by the unsettling updates, we would like to offer a few words about when, as well as when not, one *should* find dilation or sure loss unsettling. It seems to us that the attitude toward these phenomena should depend on the way the underlying probability model is interpreted.

Dilation is troubling when the set of probabilities is used as a description of uncertain inference. If the probability interval is regarded as an approximation to some underlying true probability state, akin to a confidence or posterior interval to an estimand, knowing that the interval will surely grow wider in the posterior is indeed counterproductive since the goal of inference in most cases is to tighten the interval. But in this sense, the sure loss phenomenon may just be fine, since it is common to derive disjoint yet equally valid confidence or posterior intervals from the same sampling posterior distribution, without violating any classic rules of probabilistic calculation.

On the other hand, as explained in Section 3.2, the lower and upper probabilities can be taken as acceptable prices of a gamble. Under this interpretation, any strategy that induces sure loss is absolutely unacceptable. Yet in this case, dilation has much less to worry about, since a strictly wider interval in the posterior will simply exclude the player from engaging in the called-off bet, and does not violate coherence in a decision-theoretic sense.

With precise probabilities, to condition on an observable event is to impose a restriction to the subspace defined by that event. The conditioning event itself must be measurable with respect to the original probability space. With imprecise probabilities, not all events are measurable with respect to the imprecise probability model specified on the full joint space. A crucial way the updating rules differ from one another is how they make use of this supplied conditioning information. Therefore, for any of the updating rules to function at all, they must build within themselves a particular "mechanism" of imposing the mathematical restriction specified by the observable event, when it is not currently measurable with respect to the set of probabilities the rule aims to update, much in the same way as a sampling mechanism (Kish, 1965) or missing-data mechanism (Rubin, 1976). The fact that dilation and sure loss cannot happen under the precise probability does not necessarily render them undesirable: the quality of this inference hinges on the quality of the final action they recommend. Bringing these anomalies to light allows us to study their implications, especially those unfamiliar or unexpected, on the final action.

4. BEHAVIOR OF UPDATING RULES: SOME CHARACTERIZATIONS

This section presents theoretical results on the behavior of the three updating rules discussed in this paper. We begin with the intuitive ones and progress toward those that are perhaps surprising. Unless otherwise noted, this section assumes that $\underline{P}$ is a Choquet capacity of order 2 on Ω, and $\Pi = \{P \in \mathscr{M} : P \geq \underline{P}\}$, the set of probabilities compatible with $\underline{P}$. Recall $\underline{P}_{\mathfrak{B}}$, $\underline{P}_{\mathfrak{D}}$ and $\underline{P}_{\mathfrak{G}}$ are the conditional lower probability functions according to the generalized Bayes (Definition 2.5), Dempster's (Definition 2.6) and the Geometric rules (Definition 2.7), respectively.

4.1 Generalized Bayes Rule Cannot Contract Nor Induce Sure Loss

LEMMA 4.1. *Let $\mathcal{B} = \{B_1, B_2, \ldots\}$ be a measurable and denumerable partition of Ω. For any $A \in \mathscr{B}(\Omega)$, we have*

$$\inf_{B_i \in \mathcal{B}} \underline{P}_{\mathfrak{B}}(A \mid B_i) \leq \underline{P}(A), \quad \text{and}$$

$$\sup_{B_i \in \mathcal{B}} \overline{P}_{\mathfrak{B}}(A \mid B_i) \geq \overline{P}(A).$$

PROOF. We prove by contradiction. Assume that $\inf_{B_i \in \mathcal{B}} \underline{P}_{\mathfrak{B}}(A \mid B_i) > \underline{P}(A)$. For the given A, because Π is a closed set, there exists a $P^{(A)} \in \Pi$ such that $P^{(A)}(A) = \underline{P}(A)$. The superscript notation reminds us that this probability measure can vary with the choice of A. This however does not affect the validity of applying the total probability law under this chosen $P^{(A)}$, which leads to

$$\begin{aligned}
\underline{P}(A) &= P^{(A)}(A) \\
&= \sum_{i=1}^{\infty} P^{(A)}(A \mid B_i) P^{(A)}(B_i) \\
&\geq \sum_{i=1}^{\infty} \underline{P}_{\mathfrak{B}}(A \mid B_i) P^{(A)}(B_i) \\
&\geq \sum_{i=1}^{\infty} \inf_{B_i} \underline{P}_{\mathfrak{B}}(A \mid B_i) P^{(A)}(B_i) \\
&> \sum_{i=1}^{\infty} \underline{P}(A) P^{(A)}(B_i) = \underline{P}(A),
\end{aligned}$$

resulting in a contradiction. The same argument applies to the upper probability of A. If $\sup_{B_i \in \mathcal{B}} \overline{P}_{\mathfrak{B}}(A \mid B_i) < \overline{P}(A)$, then using $\overline{P}(A) = \tilde{P}^{(A)}(A)$,

$$\begin{aligned}
\overline{P}(A) &\leq \sum_{i=1}^{\infty} \overline{P}_{\mathfrak{B}}(A \mid B_i) \tilde{P}^{(A)}(B_i) \\
&< \sum_{i=1}^{\infty} \overline{P}(A) \tilde{P}^{(A)}(B_i) = \overline{P}(A),
\end{aligned}$$

and hence again a contradiction. □

A direct consequence of Lemma 4.1 is the following thorem.

THEOREM 4.2. *Let $\mathcal{B}$ be a denumerable and measurable partition of Ω, and Π be the set of probability measures compatible with $\underline{P}$. For any event $A \in \mathscr{B}(\Omega)$, under the generalized Bayes rule*:

- *$\mathcal{B}$ cannot induce sure loss in A,*
- *$\mathcal{B}$ cannot contract A.*

The first part of Theorem 4.2, that the generalized Bayes rule avoids sure loss, is well known in the literature and is the very reason that many authors such as Walley (1991) and Jaffray (1992) consider it to be the sole choice as coherent updating rule, or the "conditioning proper." However, as we will show next, the generalized Bayes rule is also the most prone to dilation.

4.2 Generalized Bayes Rule Dilates More

LEMMA 4.3 (Generalized Bayes rule produces the widest intervals). *For all $A, B \in \mathscr{B}(\Omega)$ such that the following quantities are defined, we have*

$$\begin{aligned}
\underline{P}_{\mathfrak{B}}(A \mid B) \leq \underline{P}_{\mathfrak{D}}(A \mid B) \leq \overline{P}_{\mathfrak{D}}(A \mid B) \\
\leq \overline{P}_{\mathfrak{B}}(A \mid B)
\end{aligned} \tag{4.1}$$

and

$$\begin{aligned}
\underline{P}_{\mathfrak{B}}(A \mid B) \leq \underline{P}_{\mathfrak{G}}(A \mid B) \leq \overline{P}_{\mathfrak{G}}(A \mid B) \\
\leq \overline{P}_{\mathfrak{B}}(A \mid B).
\end{aligned} \tag{4.2}$$

That is, the conditional probability intervals resulting from Dempster's rule and the Geometric rule are always contained within those of the generalized Bayes rule. The fact that Dempster's rule produces shorter posterior intervals than that of the generalized Bayesian rule was discussed in Dempster (1967) and Kyburg (1987). Here is a simple proof that applies to both sharper rules.

PROOF. For Dempster's rule, the conditional plausibility function satisfies

$$\begin{aligned}
\overline{P}_{\mathfrak{D}}(A \mid B) &= \frac{\sup_{P \in \Pi} P(A \cap B)}{\sup_{P \in \Pi} P(B)} \leq \sup_{P \in \Pi} \frac{P(A \cap B)}{P(B)} \\
&= \overline{P}_{\mathfrak{B}}(A \mid B)
\end{aligned}$$

and by conjugacy, also $\underline{P}_{\mathfrak{D}}(A \mid B) \geq \underline{P}_{\mathfrak{B}}(A \mid B)$. Similarly for the Geometric rule, the conditional lower probability function satisfies

$$\begin{aligned}
\underline{P}_{\mathfrak{G}}(A \mid B) &= \frac{\inf_{P \in \Pi} P(A \cap B)}{\inf_{P \in \Pi} P(B)} \geq \inf_{P \in \Pi} \frac{P(A \cap B)}{P(B)} \\
&= \underline{P}_{\mathfrak{B}}(A \mid B)
\end{aligned}$$

and by conjugacy, also $\overline{P}_{\mathfrak{G}}(A \mid B) \leq \overline{P}_{\mathfrak{B}}(A \mid B)$. □

THEOREM 4.4 (Generalized Bayes rule dilates more). *Let $B \in \mathscr{B}(\Omega)$ be such that $\underline{P}(B) > 0$. Denote sets of posterior probability measures $\Pi_{\mathfrak{B}} = \{P : P \geq \underline{P}_{\mathfrak{B}}(\cdot \mid B)\}$, $\Pi_{\mathfrak{D}} = \{P : P \geq \underline{P}_{\mathfrak{D}}(\cdot \mid B)\}$ and $\Pi_{\mathfrak{G}} = \{P : P \geq \underline{P}_{\mathfrak{G}}(\cdot \mid B)\}$. Then*

$$\Pi_{\mathfrak{G}} \subseteq \Pi_{\mathfrak{B}} \quad \text{and} \quad \Pi_{\mathfrak{D}} \subseteq \Pi_{\mathfrak{B}}. \tag{4.3}$$

Theorem 4.4 is a direct consequence of Lemma 4.3, noting that $\Pi_{\mathfrak{G}}$, $\Pi_{\mathfrak{B}}$ and $\Pi_{\mathfrak{D}}$ are all convex and closed. Two more consequences of Lemma 4.3 are stated below, of which Examples 3 and 5 are respective embodiments.

COROLLARY 4.5. *If $\mathcal{B}$ incurs sure loss in A under Dempster's rule and sure gain under the Geometric rule, or vice versa, then $\mathcal{B}$ strictly dilates A under generalized Bayesian rule.*

COROLLARY 4.6. *If $\mathcal{B}$ (strictly) dilates A under either Dempster's rule or the Geometric rule, then $\mathcal{B}$ (strictly) dilates A under generalized Bayesian rule.*

Theorem 2.1 of Seidenfeld and Wasserman (1993) stated that, if dilation occurs with the generalized Bayes rule, the associated set of probabilities Π has a nonempty intersection with that of the independence plane between A and B. Thus following Corollary 4.6, we have the following.

COROLLARY 4.7. *If $\mathcal{B} = \{B, B^c\}$ dilates A under either Dempster's rule or the Geometric rule, then there exists $P^* \geq \underline{P}$ such that*

$$P^*(A \cap B) = P^*(A)P^*(B). \tag{4.4}$$

That is, dilation of an event by a binary partition under either Dempster's or the Geometric rules is a necessary condition for the posited set of probabilities to postulate event independence, since posterior intervals under both rules are contained within the generalized Bayes posterior interval.

4.3 Generalized Bayes Rule and Geometric Rule Cannot Sharpen Vacuous Prior Intervals

THEOREM 4.8 (Sharpening of vacuous intervals). *Let $\underline{P}$ be such that for the event $A \in \mathscr{B}(\Omega)$, $\underline{P}(A) = 0$, $\overline{P}(A) = 1$. For any $B \in \mathscr{B}(\Omega)$ such that $\underline{P}(B) > 0$, we have*

$$\underline{P}_{\mathfrak{G}}(A \mid B) = 0, \quad \overline{P}_{\mathfrak{G}}(A \mid B) = 1 \tag{4.5}$$

and

$$\underline{P}_{\mathfrak{B}}(A \mid B) = 0, \quad \overline{P}_{\mathfrak{B}}(A \mid B) = 1. \tag{4.6}$$

PROOF. If $\underline{P}(A) = 0$ and $\overline{P}(A) = 1$, then $\underline{P}(A \cap B) = \underline{P}(A^c \cap B) = 0$ for any B. Therefore, by (2.9) we have

$$\underline{P}_{\mathfrak{G}}(A \mid B) = \underline{P}(A \cap B)/\underline{P}(B) = 0$$

and $\overline{P}_{\mathfrak{G}}(A \mid B) = 1 - \underline{P}_{\mathfrak{G}}(A^c \mid B) = 1$, provided that the denominator is greater than zero. Furthermore, by (4.1) we have $\underline{P}_{\mathfrak{B}}(A \mid B) \leq \underline{P}_{\mathfrak{G}}(A \mid B) = 0$ and $\overline{P}_{\mathfrak{B}}(A \mid B) \geq \overline{P}_{\mathfrak{G}}(A \mid B) = 1$. □

The liberty to express partially lacking, and vacuous, prior knowledge is a prized advantage of imprecise probability over their precise, or full Bayesian, counterparts. Theorem 4.8 shows that both the generalized Bayes rule and Geometric rule are incapable of revising a vacuous prior interval to something informative for any possible outcome in the event space, whereas Dempster's rule is capable of such revision, with Example 1 being an instance. This again highlights the nonnegligible influence imposed by the rule itself, as well as the difficulty to deliver all desirable properties in one single rule. Avoiding sure loss and being able to update from complete ignorance both seem to be rather basic requirements, but to insist on both is sufficient to eliminate all three rules studied here. The following result perhaps is even more disturbing, because it says that in the world of imprecise probabilities, not only must we live with imperfections, but also accept intrinsic contradictions.

4.4 The Counteractions of Dempster's Rule and Geometric Rule

THEOREM 4.9. *If $\mathcal{B} = \{B, B^c\}$ dilates A under the Geometric rule, then it must contract A under Dempster's rule. Similarly, if $\mathcal{B}$ dilates A under Dempster's rule, then it must contract A under the Geometric rule. In both cases, the contraction is strict if the corresponding dilation is strict.*

PROOF. If $\mathcal{B}$ strictly dilates A under the Geometric rule, then for either $Z \in \mathcal{B}$

$$\underline{P}_{\mathfrak{G}}(A \mid Z) = \frac{\underline{P}(A \cap Z)}{\underline{P}(Z)} < \underline{P}(A), \tag{4.7}$$

$$\overline{P}_{\mathfrak{G}}(A \mid Z) = \frac{\overline{P}(A \cup Z^c) - \overline{P}(Z^c)}{\underline{P}(Z)} > \overline{P}(A). \tag{4.8}$$

It follows then

$$\begin{aligned}
\frac{\overline{P}_{\mathfrak{D}}(A \mid B)}{\overline{P}(A)} &= \frac{\overline{P}(A \cap B)}{\overline{P}(A) \cdot \overline{P}(B)} \\
&= \frac{\overline{P}(A \cap B)}{\overline{P}(A) \cdot (1 - \underline{P}(B^c))} \\
&< \frac{\overline{P}(A \cap B)}{\overline{P}(A) \cdot [1 - (\overline{P}(A \cup B) - \overline{P}(B))/\overline{P}(A)]} \\
&= \frac{\overline{P}(A \cap B)}{\overline{P}(A) + \overline{P}(B) - \overline{P}(A \cup B)} \leq 1,
\end{aligned}$$

where the first inequality follows from (4.8) with $Z = B^c$, and the second inequality is based on the 2-alternating nature of $\overline{P}$. (The 2-alternating nature was also implicitly used in the first inequality to ensure $\overline{P}(A \cup B) - \overline{P}(B) <$

$\overline{P}(A)$, hence the positivity of the denominator after replacing $\underline{P}(B^c)$ with an upper bound.) In a similar vein,

$$\begin{aligned}\frac{\underline{P}_{\mathfrak{D}}(A \mid B)}{\underline{P}(A)} &= \frac{\overline{P}(B) - \overline{P}(A^c \cap B)}{\overline{P}(B) \cdot \underline{P}(A)} \\ &= \frac{\underline{P}(A \cup B^c) - \underline{P}(B^c)}{\overline{P}(B) \cdot \underline{P}(A)} \\ &\geq \frac{\underline{P}(A) - \underline{P}(A \cap B^c)}{(1 - \underline{P}(B^c)) \cdot \underline{P}(A)} \\ &= \frac{\underline{P}(A) - \underline{P}(A \cap B^c)}{\underline{P}(A) - \underline{P}(B^c) \cdot \underline{P}(A)} > 1,\end{aligned}$$

where the first inequality uses the 2-monotone nature of $\underline{P}$ and the second inequality is based on (4.7) with $Z = B$. Thus we have $\overline{P}_{\mathfrak{D}}(A \mid B) < \overline{P}(A)$ and $\underline{P}_{\mathfrak{D}}(A \mid B) > \overline{P}(A)$, and clearly both inequalities still hold when we replace B by B^c because (4.7)–(4.8) hold for both $Z = B$ and $Z = B^c$. Consequently, $\mathcal{B}$ strictly contracts A under Dempster's rule. If $\mathcal{B}$ dilates A under the Geometric rule but not strictly, the inequality in either (4.7) or (4.8), but not both, may hold with equality, hence $\mathcal{B}$ contracts A under Dempster's rule but not strictly. This completes the proof for the first half of the statement.

For the second half, when $\mathcal{B}$ strictly dilates A under Dempster's rule, we have for any $Z \in \mathcal{B}$,

$$\begin{aligned}\underline{P}_{\mathfrak{D}}(A \mid Z) &= \frac{\overline{P}(A \cap Z)}{\overline{P}(Z)} > \overline{P}(A), \\ \overline{P}_{\mathfrak{D}}(A \mid Z) &= \frac{\underline{P}(A \cup Z^c) - \underline{P}(Z^c)}{\overline{P}(Z)} < \underline{P}(A).\end{aligned}$$

Noting both inequalities hold for Z and Z^c, we have

$$1 > \frac{\underline{P}(A \cup Z) - \underline{P}(Z)}{\underline{P}(A) \cdot \overline{P}(Z^c)} \geq \frac{\underline{P}(A) - \underline{P}(A \cap Z)}{\underline{P}(A) - \underline{P}(A) \cdot \underline{P}(Z)}.$$

Hence $\underline{P}(A) < \underline{P}(A \cap Z)/\underline{P}(Z) = \underline{P}_{\mathfrak{G}}(A \mid Z)$. On the other hand,

$$1 < \frac{\overline{P}(A \cap Z^c)}{\overline{P}(A) \cdot \overline{P}(Z^c)} \leq \frac{\overline{P}(A) - (\overline{P}(A \cup Z^c) - \overline{P}(Z^c))}{\overline{P}(A) - \overline{P}(A) \cdot \underline{P}(Z)}.$$

Hence $\overline{P}(A) > (\overline{P}(A \cup Z^c) - \overline{P}(Z^c))/\underline{P}(Z) = \overline{P}_{\mathfrak{G}}(A \mid Z)$. The same argument applies that if $\mathcal{B}$ dilates A under Dempster's rule but not strictly, it contracts A under the Geometric rule but not strictly. This completes the proof for the second half of the statement. □

4.5 Visualizing Relationships and Complications

EXAMPLE 5 (Pre-election poll). Suppose that we intend to study the voter intention prior to the 2016 US election. For simplicity, assume there are only two parties, represented respectively by Clinton and Trump, with one to be elected. The preelection poll consists of two questions:

TABLE 3
Hypothetical data from a voter poll consisting of two questions

Q_1	C	T	C	T	C	T	(n/a)	(n/a)	(n/a)
Q_2	Dem	Dem	Rep	Rep	(n/a)	(n/a)	Dem	Rep	(n/a)
$m(\cdot)$				$0.1 - \epsilon$					$0.2 + 8\epsilon$

1. Do you intend to vote for Trump or Clinton?
2. Do you identify more as a Republican or a Democrat?

Among all surveyed individuals, some answered both questions, some only one, and the rest did not respond. Let $Q_1 = \{\text{Trump, Clinton}\}$ denote votes for Trump and Clinton, respectively, and $Q_2 = \{\text{Republican, Democrat}\}$ denote identification with the Republican and Democratic parties, respectively. If all the percentages of response patterns are fully known, this model can be represented as a belief function. Assume the mass function $m(\cdot)$ reflecting the coarsened sampling distribution for these set-valued observations appears as Table 3 (of course, the numbers are for illustrations only).

A "tuning parameter" $\epsilon \in [-0.025, 0.1]$ is installed to create a family of mass function specifications in order to investigate the differential behavior among updating rules as a function of the coarseness of the data. The smaller the ϵ, the more the mass function concentrates on the precise observations (more survey questions answered). The larger the ϵ, the closer the random set approaches the vacuous belief function. As a function of ϵ, the prior lower and upper probabilities for Clinton are

$$\underline{P}(\mathrm{C}) = 0.3 - 3\epsilon, \quad \overline{P}(\mathrm{C}) = 0.7 + 3\epsilon.$$

The prior lower and upper probabilities for Trump, as well as for identification of either parties are numerically identical to the above, since the setup is fully symmetric with respect to both voting intention and partisanship. For example, when $\epsilon = 0$, the table above shows that altogether 40% of the respondents diligently answered both questions, 20% only identified prior partisanship, 20% only expressed current voting intentions, and another 20% did not respond at all. Thus, $m(\cdot)$ determines a pair of belief and plausibility functions which bounds the vote share for both Clinton and Trump to be within 30% and 70%.

How will information on partisanship affect the knowledge on voting intention? According to the three updating rules, the lower and upper probabilities for Clinton conditional on either values of partisanship Q_2, are as follows:

$$\begin{aligned}\underline{P}_{\mathfrak{B}}(\mathrm{C} \mid Q_2) &= \frac{0.1 - \epsilon}{0.6 + 4\epsilon}, \qquad \overline{P}_{\mathfrak{B}}(\mathrm{C} \mid Q_2) = \frac{0.5 + 5\epsilon}{0.6 + 4\epsilon}, \\ \underline{P}_{\mathfrak{D}}(\mathrm{C} \mid Q_2) &= \frac{0.2 - 2\epsilon}{0.7 + 3\epsilon}, \qquad \overline{P}_{\mathfrak{D}}(\mathrm{C} \mid Q_2) = \frac{0.5 + 5\epsilon}{0.7 + 3\epsilon}, \\ \underline{P}_{\mathfrak{G}}(\mathrm{C} \mid Q_2) &= \frac{1}{3}, \qquad \overline{P}_{\mathfrak{G}}(\mathrm{C} \mid Q_2) = \frac{2}{3}.\end{aligned}$$

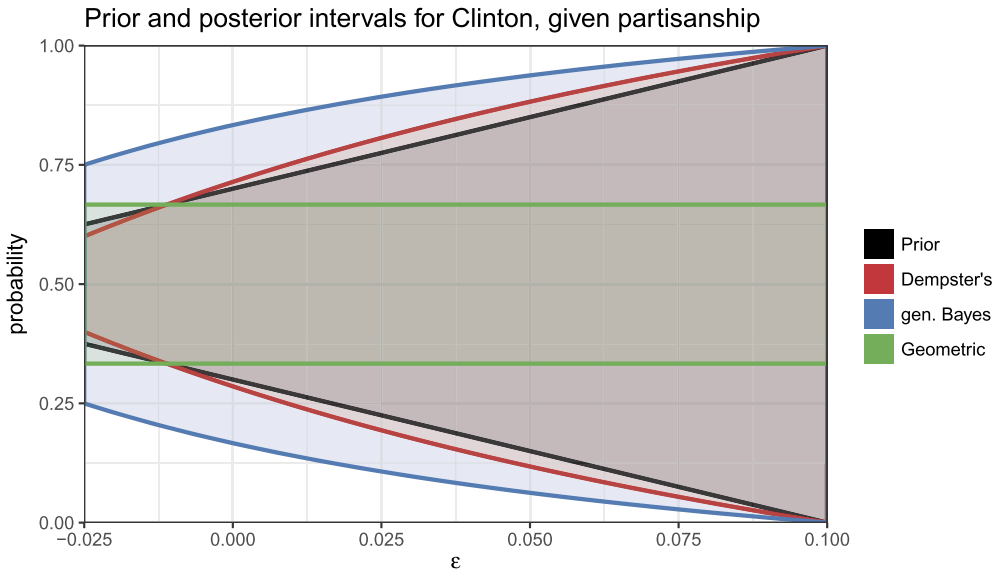

FIG. 2. *Prior probability interval for Clinton's voter support* (*black*) *and posterior probability intervals given reported partisanship according to the three updating rules* (*blue*: *generalized Bayes*, *red*: *Dempster's*, *green*: *Geometric*). *Due to full symmetry of the setup*, *contraction happens under an updating rule whenever the corresponding posterior interval depicted is contained within the prior interval*; *vice versa for dilation*.

See Figure 2 for the above quantities as functions of ϵ. We observe that:

- Under *the generalized Bayes rule*, knowledge about partisanship strictly dilates voting intention for either candidate for all $\epsilon < 0.1$. That is to say, learning the prior partisanship of an individual dilates our inference of her current voting intention, and vice versa, and this is true no matter which party or candidate is said to be favored;
- Under *Dempster's rule*, partisanship strictly dilates voting intention for either candidate for $-0.011 < \epsilon < 0.1$, and strictly contracts both for $-0.025 < \epsilon < -0.011$;
- Under *the Geometric rule*, partisanship strictly dilates voting intention for either candidate for $-0.025 < \epsilon < -0.011$, and strictly contracts both for $-0.011 < \epsilon < 0.1$. Moreover, the absolute value of the lower and upper posterior probability remained constant regardless of the value of ϵ.

Furthermore, we observe some of the phenomena discussed previously in this section. For example, the extent of dilation exhibited by the generalized Bayes rule is to a strictly larger extent than that of both Dempster's rule and the Geometric rule, if either of them does dilate. The dilation-contraction status of Dempster's rule and the Geometric rule are in full opposition to each other, switching precisely at $\epsilon = -0.011$.

5. SIMPSON'S PARADOX: AN IMPRECISE MODEL WITH AGGREGATION SURE LOSS

One may well think that all examples discussed so far lie on the boundary, if not outside, of the realm of mainstream statistical modeling. Imprecise models are not the kind of thing one just stumbles upon, they exist by intentional construction. We argue that such is not the case, that all precise models are really just the tip of an "imprecise model iceberg." Every precise model is a fully specified margin nested within a larger, ever-augmentable model, with extended features not allowed to enter the scene as the modeler lacks the knowledge to do so precisely.

Here is a concrete way to induce an imprecise model from a precise one. Take a precise model with the state space $(X_1, \ldots, X_p)$ that merits a known multivariate distribution. If we expand the model to include a previously unobservable margin X_{p+1}, the state space becomes $(p+1)$-dimensional, and the augmented model becomes imprecise. As many as $2^p - 1$ new marginal relationships—between X_{p+1} and any nonempty subset of $(X_1, \ldots, X_p)$—are left to be specified or learned. In the regression setting where a multivariate Normal model is assumed for the previous p variables, one seemingly straightforward way is to model $(X_1, \ldots, X_p, X_{p+1})$ as jointly Normal. This is a very strong assumption that takes care of all the new joint relationships. Even under such drastic simplification, the new mean and the new bivariate covariances are still left to specify, resulting in a family of $(p+1)$-dimensional Normal models.

In reality, the relationship between the existing state space and a new margin is often something about which the analyst is neither knowledgeable nor comfortable making assumptions. This is the case in observational studies, where X_{p+1} is a lurking variable which may have strong collinearity with subsets of the observed variables $(X_1, \ldots, X_p)$. Using the language of imprecise probability, we now turn to decipher Simpson's paradox, a famous and familiar setting with its far-reaching significance. The

occurrence of Simpson's paradox is proof that we have employed, likely due to lack of control, an aggregation rule that has incurred sure loss in inference.

EXAMPLE 4 CONT. (Simpson's paradox). Following the setup in Section 1, Simpson's paradox refers to an apparent contradiction between an inference on treatment efficacy at an aggregated level, $\bar{p}_{\text{obs}} < \bar{q}_{\text{obs}}$, and the inference at the disaggregated level when the covariate type of the patient has been accounted for: $p_k > q_k$ for all $k = 1, \ldots, K$. Indeed, how can a treatment be superior than its alternative in every possible way, yet be inferior *overall*?

5.1 Explicating the Aggregation Rules Underlying the Simpson's Paradox

Denote for $k = 1, \ldots, K$,

$$u_k = P(U = k \mid Z = 1), \quad v_k = P(U = k \mid Z = 0).$$

Here, $\mathbf{u}$ and $\mathbf{v}$ reflect the demographic distribution of the populations receiving the experimental and control treatments, respectively. By the law of total probability,

$$\bar{p} = \mathbf{p}^\top \mathbf{u} \quad \text{and} \quad \bar{q} = \mathbf{q}^\top \mathbf{v}, \tag{5.1}$$

thus given fixed $\mathbf{p}$ and $\mathbf{q}$, $\bar{p}$ and $\bar{q}$ are functions of $\mathbf{u}$ and $\mathbf{v}$, respectively. The marginal probabilities $\bar{p}$ and $\bar{q}$ are meant to describe an event under conditions of inferential interest, in this case, patient recovery within the two treatment arms. We refer to $\mathbf{u}$ and $\mathbf{v}$ as *aggregation rules*, functions that map conditional probabilities to a marginal probability. Aggregation rules point in reverse direction as do *updating rules* as discussed in the previous sections, which are maps from a marginal probability to a set of conditional probabilities.

Typically, measurements between different conditions are made for the purpose of a comparison, such as the evaluation of an causal effect of treatment Z on outcome Y. A comparison between $\bar{p}$ and $\bar{q}$ is *fair* if and only if the aggregation rules they employ are identical, that is, $\mathbf{u} = \mathbf{v}$ as in (5.1). This is what it means to say the comparison has been made between apples and apples. Such is the case if no confounding exists between the covariate U and the propensity of assignment, that is, $U \perp Z$.

Clearly, when $\mathbf{u} = \mathbf{v}$, $\bar{p} > \bar{q}$ if $p_k > q_k$ for all k. Hence Simpson's paradox is mathematically impossible within a fair comparison. However, for a given observed pair $\bar{p}_{\text{obs}}$ and $\bar{q}_{\text{obs}}$, have we been careful enough to enforce the *de facto* aggregation rules to equal the ideal one? That is, do we have that the observed comparison is *fair enough*, that is, a common rule $\mathbf{v}$ such that approximately,

$$\mathbf{u}_{\text{obs}} \doteq \mathbf{v} \quad \text{and} \quad \mathbf{v}_{\text{obs}} \doteq \mathbf{v}? \tag{5.2}$$

For certain values of $\mathbf{p}$ and $\mathbf{q}$, it is entirely possible that suitable realizations of $(\mathbf{u}_{\text{obs}}, \mathbf{v}_{\text{obs}})$ could result in $\bar{p}_{\text{obs}} < \bar{q}_{\text{obs}}$. To be exact, these are $\mathbf{p}$ and $\mathbf{q}$ values satisfying $\max_k q_k > \min_k p_k$. At least one, and possibly both realizations of $\mathbf{u}_{\text{obs}}$ and $\mathbf{v}_{\text{obs}}$ play differentially to the relative weaknesses of $\mathbf{p}$, that is, coordinates of smaller magnitude, and the strengths of $\mathbf{q}$ accordingly. When this preferential weighting, also known as *confounding*, is strong enough to reverse the perceived stochastic dominance of the outcome variable under either treatment, an apparent paradox is induced. Randomization procedures effectively put quality guarantees on the fairness of comparison; as the sample size n grows larger, (5.2) holds with high probability with deviations quantifiable with respect to $\mathbf{p}$ and $\mathbf{q}$ that is immune against all U, observed or unobserved.

5.2 The Paradox Is Sure Loss

Simpson's paradox is reminiscent of the "sure loss" phenomenon we saw in earlier sections. Indeed, when not conditioned on U, if asked to pick a bet between the experimental and control treatments, we would prefer the control treatment over the experimental one. But once conditioned on U, the experimental treatment suddenly became the superior bet regardless of U's value. One is thus set to surely lose money by engaging in a combination of these two bets. This is formalized by the following theorem, where $\mathcal{S}_K$ is the standard K-simplex defined by $\{(v_1, \ldots, v_K) : \sum_{k=1}^K v_k = 1; v_k \geq 0, k = 1, \ldots, K\}$.

THEOREM 5.1 (Equivalence of Simpson's paradox and aggregation sure loss). *Let Λ be a convex hull in $[0,1]^K$ characterized by the pair of elementwise upper and lower bounds $(\mathbf{p}, \mathbf{q})$. That is,*

$$\Lambda = \{\boldsymbol{\lambda} \in [0,1]^K : q_k \leq \lambda_k \leq p_k, k = 1, \ldots, K\}.$$

Let $\mathcal{V} \subseteq \mathcal{S}_K$ be a closed set of aggregation rules, *and $\mathbf{u} \in \mathcal{S}_K$. Then $\mathbf{u}$ incurs sure loss on Λ relative to $\mathcal{V}$ if and only if $(\mathbf{u}, \mathbf{v})$ induces Simpson's paradox in $(\mathbf{p}, \mathbf{q})$ for all $\mathbf{v} \in \mathcal{V}$.*

PROOF. Denote the set of marginal probability derived from Λ under the set of aggregation rules $\mathcal{V}$ as $\mathcal{P}_\mathcal{V} = \{\boldsymbol{\lambda}^\top \mathbf{v} : \boldsymbol{\lambda} \in \Lambda, \mathbf{v} \in \mathcal{V}\}$. By the closeness of both Λ and $\mathcal{V}$, we have

$$\inf \mathcal{P}_\mathcal{V} = \inf_{\mathbf{v} \in \mathcal{V}} \mathbf{q}^\top \mathbf{v} \quad \text{and} \quad \sup \mathcal{P}_\mathcal{V} = \sup_{\mathbf{v} \in \mathcal{V}} \mathbf{p}^\top \mathbf{v}, \tag{5.3}$$

and

$$\mathbf{p}^\top \mathbf{u} = \sup_{\boldsymbol{\lambda} \in \Lambda} \boldsymbol{\lambda}^\top \mathbf{u} \quad \text{and} \quad \mathbf{q}^\top \mathbf{u} = \inf_{\boldsymbol{\lambda} \in \Lambda} \boldsymbol{\lambda}^\top \mathbf{u}. \tag{5.4}$$

Employing Definition 3.3, to say that $\mathbf{u}$ incurs sure loss on Λ relative to $\mathcal{V}$ means that

$$\sup_{\boldsymbol{\lambda} \in \Lambda} \boldsymbol{\lambda}^\top \mathbf{u} < \inf \mathcal{P}_\mathcal{V} \quad \text{or} \quad \inf_{\boldsymbol{\lambda} \in \Lambda} \boldsymbol{\lambda}^\top \mathbf{u} > \sup \mathcal{P}_\mathcal{V}. \tag{5.5}$$

On the other hand, to say that for every $\mathbf{v} \in \mathcal{V}$, $(\mathbf{u}, \mathbf{v})$ induces Simpson's paradox in $(\mathbf{p}, \mathbf{q})$ means that

$$\mathbf{p}^\top \mathbf{u} < \inf_{\mathbf{v} \in \mathcal{V}} \mathbf{q}^\top \mathbf{v} \quad \text{or} \quad \mathbf{q}^\top \mathbf{u} > \sup_{\mathbf{v} \in \mathcal{V}} \mathbf{p}^\top \mathbf{v}. \tag{5.6}$$

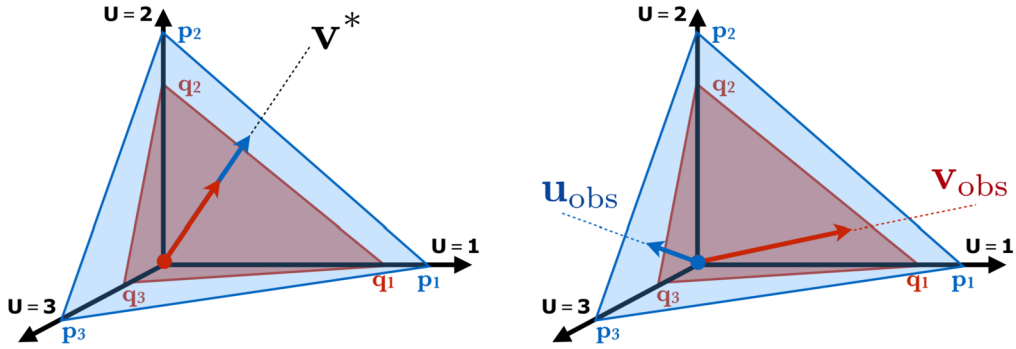

FIG. 3. *Ideal aggregating rules guarantee the comparison between treatment arms is made on a fair ground. Observed Simpson's paradox is strong evidence that the de facto aggregating rules are fair for comparison. Left*: *if* $p_k > q_k$ *for all* k, *then* $\mathbf{p}^\top \mathbf{v} > \mathbf{q}^\top \mathbf{v}$ *for all* $\mathbf{v}$; *Right*: *disparate* $\mathbf{u}_{obs}$ *and* $\mathbf{v}_{obs}$ *make possible* $p_{obs} < q_{obs}$. *Note that* Π *in Theorem* 5.1 *is the convex hull sandwiched between the blue* ($\mathbf{p}$) *and red* ($\mathbf{q}$) *hyperplanes in the first octant.*

Identities (5.3)–(5.4) trivially imply the equivalence between (5.5) and (5.6). □

We remark that, in Definition 3.3, sure loss is defined with respect to a single conditioning rule because the prior/marginal lower and upper probabilities $\underline{P}$ and $\overline{P}$ are treated as given. Such is not the case with the sure loss concept in Theorem 5.1. We must first define $\mathcal{V}$, a set of aggregation rules deemed desirable for the purpose of the study. $\mathcal{V}$ implies a prior/marginal probability interval, only relative to which the behavior of the other aggregation rule $\mathbf{u}$ can be discussed. One can check that the relationship between $\mathbf{u}$ and $\mathbf{v}$ is reciprocal, that is, if $\mathbf{u}$ induces sure loss relative to $\mathbf{v}$, then $\mathbf{v}$ induces sure loss relative to $\mathbf{u}$. Thus, we can talk about an *aggregation scheme* as an ordered pair of rules $(\mathbf{u}, \mathbf{v})$, and its characteristics as whether it incurs sure loss relative to itself, whether it induces the paradox in $(\mathbf{p}, \mathbf{q})$, and so on.

A connection between Simpson's paradox and the *atomic lower and upper probability* (ALUP) model of Herron, Seidenfeld and Wasserman (1997) is made below. A set of probabilities $\Pi_{(\mathbf{p},\mathbf{q})}$ is an ALUP generated by $(\mathbf{p}, \mathbf{q}) \in [0, 1]^{2K}$, if

$$\Pi_{(\mathbf{p},\mathbf{q})} = \{\boldsymbol{\pi} \in \mathcal{S}_K : \sup \pi_k = p_k, \inf \pi_k = q_k\} \tag{5.7}$$

LEMMA 5.2 (ALUP models). *If an aggregation scheme* $(\mathbf{u}, \mathbf{v})$ *induces Simpson's paradox in* $(\mathbf{p}, \mathbf{q})$, *it incurs sure loss relative to itself on the ALUP model* $\Pi_{(\mathbf{p},\mathbf{q})}$ *as defined in* (5.7).

PROOF. Without loss of generality, suppose an aggregation scheme $(\mathbf{u}, \mathbf{v})$ induces Simpson's paradox in $(\mathbf{p}, \mathbf{q})$ in the form of $\mathbf{p}^\top \mathbf{u} = \sup_{\boldsymbol{\lambda} \in \Lambda} \boldsymbol{\lambda}^\top \mathbf{u} < \inf_{\boldsymbol{\lambda} \in \Lambda} \boldsymbol{\lambda}^\top \mathbf{v} = \boldsymbol{q}^\top \mathbf{v}$. But since $\Pi_{(\mathbf{p},\mathbf{q})}$ is a closed and convex subset of Λ, we have $\sup_{\boldsymbol{\lambda} \in \Lambda} \boldsymbol{\lambda}^\top \mathbf{u} \geq \sup_{\boldsymbol{\pi} \in \Pi_{(p,q)}} \boldsymbol{\pi}^\top \mathbf{u}$ and $\inf_{\boldsymbol{\lambda} \in \Lambda} \boldsymbol{\lambda}^\top \mathbf{v} \leq \inf_{\boldsymbol{\pi} \in \Pi_{(p,q)}} \boldsymbol{\pi}^\top \mathbf{v}$, hence the "only if" part of Theorem 5.1 still holds. □

5.3 Implication on Inference

In Example 4, the description of the model is precise with the conditional values $\mathbf{p}$ and $\mathbf{q}$, as well as the marginal values $\bar{p}_{\text{obs}}$ and $\bar{q}_{\text{obs}}$. The model is imprecise, and in fact completely vacuous, on the aggregation rules $(\mathbf{u}_{\text{obs}}, \mathbf{v}_{\text{obs}})$ which gave rise to the observed marginal values.

In order for the observed marginal probabilities $\bar{p}_{\text{obs}}$ and $\bar{q}_{\text{obs}}$ to yield a meaningful comparison, we must have clear answers to the following two questions regarding $\mathbf{u}_{\text{obs}}$ and $\mathbf{v}_{\text{obs}}$:

1. Are they equal?
2. What is the mutual value $\mathbf{v}$ they both should be equal to?

An affirmative answer to the first question ensures that $\bar{p}_{\text{obs}}$ and $\bar{q}_{\text{obs}}$ are at least on a comparable footing. For example, for the evaluation of an causal effect of Z on Y, regardless of the population of interest, it must be ensured that no confounding between the covariate U and the propensity of assignment took place, that is, $U \perp Z$. That is why Simpson's paradox is a sanity check for any apparent causal relationship, as the paradox constitutes sufficient (but not necessary) evidence there is nonnegligible confounding between U and Z, a telltale sign that one is comparing apples with oranges.

Much classic and contemporary literature on causal inference sensitivity analysis, for example, Cornfield et al. (1959), Ding and VanderWeele (2016), hinge on establishing deterministic bounds to exclude scenarios that are in essence Simpson's paradoxes, as well as quantifying the probability of population-level paradox given observed paradox in the sample, for example, Pavlides and Perlman (2009). If the assignment Z cannot be controlled in one or both treatment arms, the aggregation rule is no longer chosen by the investigator but rather left self-selected, in all or in part by the observational mechanism. In particular, if arbitrary confounding can be present in

both treatment arms, $\mathbf{u}$ and $\mathbf{v}$ can take up any value in $\mathcal{S}^K$. It is also entirely possible that controlled randomization or weighting is available in only one of the treatment arms, or on a subset of levels of U, reflecting an aggregation rule as a mixture of intentional choice and self-selection.

It is also crucial that the ideal aggregation rule $\mathbf{v}$, the mutual value for $\mathbf{u}_{\text{obs}}$ and $\mathbf{v}_{\text{obs}}$, is a conscious choice made to reflect the scientific question of interest. Two typical situations that give rise to natural choices of $\mathbf{v}$ are:

- to infer about population average treatment effect, choose $\mathbf{v}$ to be the oracle probability distribution of patients' covariates in the population;
- to make inference about a particular patient's treatment effect, choose $\mathbf{v} = (0 \cdots 0\ 1_{(U_i=k)}\ 0 \cdots 0)^\top$, the indicator vector matching the patient's covariate value U_i with its level k.

One can devise a range of choices of $\mathbf{v}$ to reflect any amount of intermediate pooling within what is deemed as the relevant subpopulation. As discussed in Liu and Meng (2014, 2016), what defines the game of *individualized inference* is picking the $\mathbf{v}$ at the appropriate resolution level while subject to the tradeoff between population relevance and estimation robustness.

Choosing the right $\mathbf{v}$ and enforcing $\mathbf{u}_{\text{obs}} = \mathbf{v}_{\text{obs}} = \mathbf{v}$ is not merely a mathematical decision on paper, but rather entails action in a real-life observational environment, one that likely involves the physical activities of stratification and randomization such as controlled experiments and survey designs. Only through doing so can we make sure the de facto aggregation rules are equal to the ideal rule, or equivalently that we know executable ways to adjust for the differences between these quantities, for example, through retrospective weighting. Failure to acknowledge the distinction and potential differences among $\mathbf{v}$, $\mathbf{u}_{\text{obs}}$, and $\mathbf{v}_{\text{obs}}$ paves the way not only for Simpson's paradox, but also equivalently for endorsing mythical statistical aggregation rules with the potential to exhibit incoherent behavior, and the worst of all, to mislead ourselves in making the wrong treatment or policy decisions, a sure loss in a real sense.

6. FOOD FOR THOUGHT

6.1 Imprecise Models: Extended Expressions of Uncertainty

When more information is observed, we expect the variability associated with the inferential target to decrease. This property is possessed by many trustworthy Bayesian and frequentist procedures relying on precise model structures. Those that bring the most variability reduction for unit increase of observed information are praised as statistically efficient.

However, efficiency is only desirable if we are absolutely sure that information is utilized in the correct way. The ability of an efficient method to distinguish between useful and harmful variations in the data is supplied by the assumption underlying the model. These assumptions are sometimes made out of convenience, and sometimes out of the limited expressions of uncertainty that precise statistical models permit. Balch, Martin and Ferson (2019) observed the paradox of *probability dilution*: lower quality tracking data, when expressed via a sampling model with inflated variance, apparently increases the confidence in the inference that two satellites would not collide. The uncertainty about data acquisition got coerced into a precise piece of modeling assumption, which backfires and brings misleading precision in inference.

Probabilistic modeling is not all about convergence. A responsible modeler certainly would like to know if she does not actually have the right means to converge to the truth. She would like to articulate uncertainty about the state of knowledge, without conflating it with sampling variability which will go away as data accumulate. If additional data do not carry information beneficial with respect to the current state of knowledge, a truly intelligent model ought to refuse to further reduce inferential variability based on these data, such that additional data will do no harm.

Even within the realm of precise models, "doing no harm" is a requirement that can be easily violated when the model is misspecified. As demonstrated in Meng and Xie (2014), more data do not automatically lead to narrower confidence intervals even in ordinary least squares (OLS) regression. If a homogeneous variance model is applied to data with heteroskedasticity, the naturally equally weighted OLS de-facto gives observations with larger noise more weight than they deserve. The width of the confidence interval can increase, sometimes substantially, with the size of our data. Indeed, a heteroskedastic regression model without knowledge of how the heteroskedasticity arises cannot teach itself to weight a new data point without mixing signal with noise, an obvious reflection of an inherent structural deficiency in the model.

Equipped with such intuition, it becomes natural to view dilation and other anomalies with imprecise models not as annoying bugs, but rather helpful warning signs. They reflect a genuine, structural kind of uncertainty about the underlying set of probabilistic models employed. The upper and lower probability intervals, be they prior or posterior, marginal or conditional, do not merely measure the lack of information from pinning down the inferential target. They also reflect the incomplete knowledge on the modeler's part, from knowing even how to measure such lack of information. These unsettling phenomena are all symptoms when the inherent incompleteness of modeling knowledge gets in the way of learning more about the inference question. That is when

observations, which normally would bring in more information, may just become points of additional confusion, if we do not recognize their diagnostic values.

As discussed in Section 1, association characterizes how probability about one thing should change after another thing has been learned. It is the fundamental means through which observed information contribute to a model. The sign of the association gives the sense of direction, such as seen from the coefficients in regression models. The magnitude of the association implies an order of priority, such as in large scale genome-wide association studies and elsewhere where correlation coefficients are used as test statistics. Plentiful association is the indication of signal strength, potential discovery and the prospect of a causal relationship. The absence of association, on the other hand, is just as desirable when used to justify independence assumptions, creating a blanket of simplicity on which small-world models can be built and trusted. The three types of associations (positive, negative and independence) correspond to the three possible directions of change as the probability of an event updates from the prior to the posterior according to the Bayes rule. In precise probabilities, these three types of associations exhaust all possibilities of information contribution from one event to another.

Imprecise models expand the landscape of information contribution, because the probabilistic description assigned to each event is no longer singular. The upper and lower probabilities considered in this paper deliver a closed interval $[\underline{P}(A), \overline{P}(A)]$ of possibly nonnegligible width. Generalized notions of association and independence, which characterize the direction of change from prior to posterior, are yet to be defined for sets of probabilities. Phenomena like dilation, contraction and sure loss explored in this paper are hinting at novel types of information contribution, as model uncertainty revealed through them can be particularly informative and welcome. The ability to send this message is a unique and powerful feature of imprecise models, as well as those that utilize nonadditive measures (Balch, Martin and Ferson, 2019).

6.2 Assumption Incommensurability and Conditioning Protocol

As revealed in Section 3.3, each imprecise probability updating rule is constantly faced with the problem that the conditioning information may not be measurable with respect to the very imprecise probability it is trying to update. As a consequence, they each effectively build within themselves a mechanism for imposing mathematical restrictions generated by a given event B. This is why, as far as we can see, the situation in the world of imprecise probability is more confusing and clearer at the same time. It is more confusing because the notation $\underline{P}_\bullet$ and $\overline{P}_\bullet$ carry meanings contingent upon the $\bullet$-rule we choose. Yet, different rules are built upon different mechanisms for imposing the mathematical restriction specified by an event partition $\mathcal{B}$, in a much similar vein to the sampling and missing data mechanisms mentioned previously, potentially supplying a variety of options suitable for different situations that users may choose from, as long as they are well informed of the implied assumptions of each rule. In this sense, the situation is clearer, because the imprecise nature should compel the users to be explicit about the imposed mechanisms in order to proceed. Below we illustrate this point.

EXAMPLE 2 CONT. (The boxer, the wrestler and the God's coin). Recall the boxer and wrestler example in which there exists a priori, a fair coin and vacuous knowledge of the two fighters. Our analysis in Section 3 showed that upon knowing $X = Y$, Dempster's rule will judge the posterior probability of boxer's win to be precisely half, whereas generalized Bayes rule will remain that the chance is anywhere within $[0, 1]$. We realize that the witness who relayed the message $X = Y$ could have meant it in (at least) two different ways:

1. that he happened to see both the coin flip and the match between the two fighters, and the results of the two events were identical;
2. that he somehow miraculously knew that the coin toss *decides* the outcome of the match, as if the coin is God's pseudorandom number generator.

If the first meaning is taken, as most of us naturally do, it seems that the generalized Bayes answer makes sense. After all, since we do not know the relationship between two co-observed phenomenon, the worst case scenario would be to admit all possibilities, including the most extreme forms of dependence, when deriving the probability interval.

However, if the head of the coin dictates the triumph of the boxer, and the former event is known precisely as a toss-up, it makes sense to think of the match as a true toss-up as well. In this case, it is rightful to call for a transferral of the a priori precise probability of X onto the a priori vacuous Y. The same logic would apply had we been told $X \neq Y$, in the sense that the head of the coin dictates the triumph of the wrestler. In both cases, the update is akin to adding another piece of structural knowledge to the model itself.

This example reflects a point made by Shafer (1985). In order for probabilistic conditioning to be properly interpreted, it is crucial to have a "protocol" specifying what information *can* be learned, in addition to learning the actual information itself. Updating in absence of a protocol, or more dangerously under an unacknowledged, implicit protocol, can produce complications to the interpretation

of the output inference. Dilation and sure loss, phenomena exclusive to imprecise probability, are striking instances that demonstrate such danger. Discrepancies among the three updating rules reflect the different ways the same incoming message might be interpreted. Each conditioning rule effectively creates a world of alternative possible observations, hence a protocol is de facto in place, only hidden behind these explicit-looking rules.

When performing updates in the boxer and wrestler's case, the distinction between conditioning protocols underlying the solutions we have offered so far is one between *factual* versus *incidental* knowledge spaces. Knowing $X = Y$ is a possible outcome and by chance observing it constitutes incidental knowledge. Knowing that $X = Y$ is the factual state of the nature is knowledge of a fundamentally different type, one that is much more restrictive and powerful at the same time: in other words, $X \neq Y$ cannot, could not and will not happen. Unlike their incidental counterparts, claiming either $X = Y$ or $X \neq Y$ as factual necessarily makes them incommensurable with one another, even over sampling repetitions. That is to say, if either $X = Y$ or $X \neq Y$ are to be hard-coded into the model, they will each result in a model distinct from the other in a way that their respective posterior judgments about the same event, say $Y = 1$, are not meant to enter the same law of total probability calculation. If we are willing to admit either $X = Y$ or $X \neq Y$ as factual evidence to condition on, they can no longer be regarded as a partition of the full space like they did back in Section 3.1; the model must also anticipate to deal with a whole range of other possible relationships between X and Y that are nondeterministic, as part of the conditioning protocol in Shafer's sense.

The distinction between factual versus incidental knowledge updating are referred to as *revision* versus *focusing* in the imprecise probability literature, and reflect the ideologies behind the updating rules; see Smets (1991), Miranda and Montes (2015) for more on the matter. Whether a rule is applicable to a particular imprecise model would consequently depend on a judgment of knowledge type, as well as what questions we want to answer. Within a precise modeling framework, the knowledge type for conditioning is typically coded into the conditioning event itself, which might be on an enhanced probabilistic space but without increasing the resolution of the original (marginal) model because it is already at the highest possible resolution. Hence, one universal updating rule is sufficient. Under an imprecise model, such a resolution-preserving encoding may not be possible because of the low resolution nature of the original model. Various rules then have been and will be invented to carry out the update as a qualitative rescue for the model's inability to quantify the knowledge types within its original resolution. This makes the judgment of knowledge types particularly pronounced, and serves as a reminder of the precise nature of the conditioning operation in statistical learning. If the applicability and subtitles of each updating rule is not explicated, the resulting inference is subject to increased vulnerability and confusion, even leading to paradoxical phenomena such as studied in this paper.

6.3 Imprecise Probability, Precise Decisions

Seeing a myriad of sensible and nonsensible answers produced by the updating rules of imprecise models, one may wonder if anything certain, or close to certain, can be inferred from these models at all without stirring up a controversy. To this end, we discuss a final twist to the three prisoners' story.

EXAMPLE 3 CONT. (Three prisoners' Monty Hall). Having heard from the guard that B will not receive parole, prisoner A is presented with an option to switch his identity with prisoner C: that is, the next morning A will be met with the fate of C (and C that of A), both having been decided unbeknownst to them. Is this a good idea for A?

The answer is unequivocally yes. The above is a recast of the Monty Hall problem in which you, the contestant standing in front of a randomly chosen door (prisoner A), have just been shown a door with a goat behind it ("B will be executed"), and are contemplating a switch to the other unopened door (the identity of prisoner C) for a better chance of winning the new car (parole). By the calculations in (3.9), we know that under the generalized Bayes rule

$$\begin{aligned}&\overline{P}_{\mathfrak{B}}(A \text{ lives} \mid \text{guard says } B)\\ &\quad = \underline{P}_{\mathfrak{B}}(C \text{ lives} \mid \text{guard says } B),\end{aligned}$$

suggesting that a switch will under no circumstances hurt the chance of A's survival. Without switching, A's best chance of surviving does not exceed C's worst chance of living. Moreover, as the most conservative rule of all, the (almost) separation of the two generalized Bayes posterior probability intervals guarantees the same for the other updating rules as well. Therefore, the action of identity switching should be recommended to A without reservation, regardless of the choice of rule among the three discussed. (Without changing the problem setup, it is essentially disallowed for more than one prisoner to inquire with the guard, either independently or simultaneously. Thus we never have to recommend identity switching to more than one prisoner, which would otherwise create a different paradox.) The unanimity in decision is due to the (very) low resolution nature of the action space, often binary (e.g., switching or not), allowing different high-resolution probabilistic statements to admit the same low resolution classification in the action space.

ACKNOWLEDGMENTS

We thank Arthur Dempster, Haosui Duanmu, Keli Liu, Glenn Shafer, Teddy Seidenfeld and anonymous reviewers for helpful discussions and comments, and Steve Finch for careful proofreading. Research of X.-L. Meng is supported in part by the John Templeton Foundation Grant 52366, and that of R. Gong by the National Science Foundation Grant DMS-1916002.

REFERENCES

BALCH, M. S., MARTIN, R. and FERSON, S. (2019). Satellite conjunction analysis and the false confidence theorem. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **475** 20180565, 20. MR3999720 https://doi.org/10.1098/rspa.2018.0565

BILLINGSLEY, P. (2013). *Convergence of Probability Measures*. Wiley, New York.

BLYTH, C. R. (1972). On Simpson's paradox and the sure-thing principle. *J. Amer. Statist. Assoc.* **67** 364–366, 373–381. MR0314156

CORNFIELD, J., HAENSZEL, W., HAMMOND, E. C., LILIENFELD, A. M., SHIMKIN, M. B. and WYNDER, E. L. (1959). Smoking and lung cancer: Recent evidence and a discussion of some questions. *J. Natl. Cancer Inst.* **22** 173–203.

DEMPSTER, A. P. (1967). Upper and lower probabilities induced by a multivalued mapping. *Ann. Math. Stat.* **38** 325–339. MR0207001 https://doi.org/10.1214/aoms/1177698950

DIACONIS, P. (1978). Review of "A mathematical theory of evidence" (G. Shafer). *J. Amer. Statist. Assoc.* **73** 677–678.

DIACONIS, P. and ZABELL, S. (1983). Some alternatives to Bayes' Rule. Stanford University, CA. Department of Statistics.

DING, P. and VANDERWEELE, T. J. (2016). Sensitivity analysis without assumptions. *Epidemiology* **27** 368.

FAGIN, R. and HALPERN, J. Y. (1987). A new approach to updating beliefs. In *Proceedings of the Sixth Conference on Uncertainty in Artificial Intelligence* 317–325.

FYGENSON, M. (2008). Modeling and predicting extrapolated probabilities with outlooks. *Statist. Sinica* **18** 9–90. MR2416904

GELMAN, A. (2006). The boxer, the wrestler, and the coin flip: A paradox of robust Bayesian inference and belief functions. *Amer. Statist.* **60** 146–150. MR2224212 https://doi.org/10.1198/000313006X106190

GONG, R. and MENG, X. L. (2021). Probabilistic underpinning of imprecise probability for statistical learning with low-resolution information. Technical Report.

GOOD, I. (1974). A little learning can be dangerous. *British J. Philos. Sci.* **25** 340–342.

HANNIG, J. and XIE, M. (2012). A note on Dempster–Shafer recombination of confidence distributions. *Electron. J. Stat.* **6** 1943–1966. MR2988470 https://doi.org/10.1214/12-EJS734

HANNIG, J., IYER, H., LAI, R. C. S. and LEE, T. C. M. (2016). Generalized fiducial inference: A review and new results. *J. Amer. Statist. Assoc.* **111** 1346–1361. MR3561954 https://doi.org/10.1080/01621459.2016.1165102

HEITJAN, D. F. (1994). Ignorability in general incomplete-data models. *Biometrika* **81** 701–708. MR1326420 https://doi.org/10.1093/biomet/81.4.701

HEITJAN, D. F. and RUBIN, D. B. (1991). Ignorability and coarse data. *Ann. Statist.* **19** 2244–2253. MR1135174 https://doi.org/10.1214/aos/1176348396

HERRON, T. SEIDENFELD, T. and WASSERMAN, L. (1994). The extent of dilation of sets of probabilities and the asymptotics of robust Bayesian inference. In *PSA*: *Proceedings of the Biennial Meeting of the Philosophy of Science Association* 250–259.

HERRON, T., SEIDENFELD, T. and WASSERMAN, L. (1997). Divisive conditioning: Further results on dilation. *Philos. Sci.* **64** 411–444. MR1605648 https://doi.org/10.1086/392559

HUBER, P. J. and STRASSEN, V. (1973). Minimax tests and the Neyman–Pearson lemma for capacities. *Ann. Statist.* **1** 251–263. MR0356306

JAFFRAY, J.-Y. (1992). Bayesian updating and belief functions. *IEEE Trans. Syst. Man Cybern.* **22** 1144–1152. MR1202571 https://doi.org/10.1109/21.179852

KISH, L. (1965). *Survey Sampling*. Wiley, New York, NY.

KOHLAS, J. (1991). The reliability of reasoning with unreliable arguments. *Ann. Oper. Res.* **32** 67–113. MR1128173 https://doi.org/10.1007/BF02204829

KRUSE, R. and SCHWECKE, E. (1990). Specialization—a new concept for uncertainty handling with belief functions. *Int. J. Gen. Syst.* **18** 49–60.

KYBURG, H. E. (1987). Bayesian and non-Bayesian evidential updating. *Artificial Intelligence* **31** 271–293.

LIU, K. and MENG, X.-L. (2014). Comment: A fruitful resolution to Simpson's paradox via multiresolution inference. *Amer. Statist.* **68** 17–29. MR3303829 https://doi.org/10.1080/00031305.2014.876842

LIU, K. and MENG, X-L. (2016). There is individualized treatment. Why not individualized inference?. *Annu. Rev. Stat. Appl.* **3** 79–111.

MARTIN, R. and LIU, C. (2016). *Inferential Models*: *Reasoning with Uncertainty*. *Monographs on Statistics and Applied Probability* **147**. CRC Press, Boca Raton, FL. MR3618727

MENG, X.-L. and XIE, X. (2014). I got more data, my model is more refined, but my estimator is getting worse! Am I just dumb? *Econometric Rev.* **33** 218–250. MR3170847 https://doi.org/10.1080/07474938.2013.808567

MIRANDA, E. and MONTES, I. (2015). Coherent updating of non-additive measures. *Internat. J. Approx. Reason.* **56** 159–177. MR3278790 https://doi.org/10.1016/j.ijar.2014.05.003

MORGAN, J. P., CHAGANTY, N. R., DAHIYA, R. C. and DOVIAK, M. J. (1991). Let's make a deal: The player's dilemma. *Amer. Statist.* **45** 284–287.

MOSTELLER, F. (1965). *Fifty Challenging Problems in Probability with Solutions*. Courier Corporation, North Chelmsford, MA.

NGUYEN, H. T. (1978). On random sets and belief functions. *J. Math. Anal. Appl.* **65** 531–542.

PAVLIDES, M. G. and PERLMAN, M. D. (2009). How likely is Simpson's paradox? *Amer. Statist.* **63** 226–233. MR2750346 https://doi.org/10.1198/tast.2009.09007

PEARL, J. (1990). Reasoning with belief functions: An analysis of compatibility. *Internat. J. Approx. Reason.* **4** 5–6 363–389.

PEDERSEN, A. P. and WHEELER, G. (2014). Demystifying dilation. *Erkenntnis* **79** 1305–1342. MR3274419 https://doi.org/10.1007/s10670-013-9531-7

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

SCHWEDER, T. and HJORT, N. L. (2016). *Confidence*, *Likelihood*, *Probability*: *Statistical Inference with Confidence Distributions*. *Cambridge Series in Statistical and Probabilistic Mathematics* **41**. Cambridge Univ. Press, New York. MR3558738 https://doi.org/10.1017/CBO9781139046671

SEIDENFELD, T. and WASSERMAN, L. (1993). Dilation for sets of probabilities. *Ann. Statist.* **21** 1139–1154. MR1241261 https://doi.org/10.1214/aos/1176349254

SHAFER, G. (1976). *A Mathematical Theory of Evidence*. Princeton Univ. Press, Princeton, NJ.

SHAFER, G. (1979). Allocations of probability. *Ann. Probab.* **7** 827–839.

SHAFER, G. (1985). Conditional probability. International Statistical Review/Revue Internationale de Statistique 261–275.

SIMPSON, E. H. (1951). The interpretation of interaction in contingency tables. *J. Roy. Statist. Soc. Ser. B* **13** 238–241.

SMETS, P. (1991). About updating. In *Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence* 378–385.

SMETS, P. (1993). Belief functions: The disjunctive rule of combination and the generalized Bayesian theorem. *Internat. J. Approx. Reason.* **9** 1–35.

SUPPES, P. and ZANOTTI, M. (1977). On using random relations to generate upper and lower probabilities: Foundations of probability and statistics, III. *Synthese* **36** 427–440. MR0517217 https://doi.org/10.1007/BF00486106

WALLEY, P. (1981). Coherent lower (and upper) probabilities Statistics Research Report 22, University of Warwick, Coventry.

WALLEY, P. (1991). *Statistical Reasoning with Imprecise Probabilities*. Taylor & Francis, Oxford, UK.

WASSERMAN, L. A. and KADANE, J. B. (1990). Bayes' theorem for Choquet capacities. *Ann. Statist.* **18** 1328–1339. MR1062711 https://doi.org/10.1214/aos/1176347752

XIE, M. and SINGH, K. (2013). Confidence distribution, the frequentist distribution estimator of a parameter: A review. *Int. Stat. Rev.* **81** 3–39. MR3047496 https://doi.org/10.1111/insr.12000

YAGER, R. R. (1987). On the Dempster–Shafer framework and new combination rules. *Inform. Sci.* **41** 93–137.

YAGER, R. R. and LIU, L., eds. (2008). *Classic Works of the Dempster–Shafer Theory of Belief Functions* **219**. Springer, New York, NY.