Statistical Science
0, Vol. 0, No. 00, 1–22
https://doi.org/10.1214/19-STS765

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

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:

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

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).

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 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 $k = 1, \ldots, K$,

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

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,

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

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 $\overline{P}(B) > 0$,

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

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

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 $\underline{P}(B) > 0$,

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

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

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

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

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*

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

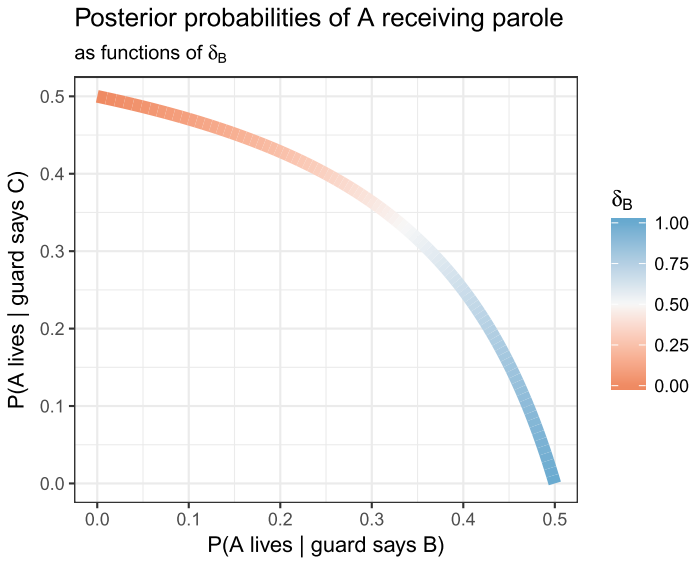

FIG. 1. *Posterior probabilities of prisoner A receiving parole given the guard's two possible answers, as a function of the guard's reporting bias* δ_B (3.7).

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*

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

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 $Z \in \mathcal{B}$,

$$\underline{P}_{\mathfrak{D}}(A \mid Z) = \frac{\overline{P}(A \cap Z)}{\overline{P}(Z)} > \overline{P}(A),$$

$$\overline{P}_{\mathfrak{D}}(A \mid Z) = \frac{\underline{P}(A \cup Z^c) - \underline{P}(Z^c)}{\overline{P}(Z)} < \underline{P}(A).$$

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:

$$\underline{P}_{\mathfrak{B}}(\mathrm{C} \mid Q_2) = \frac{0.1 - \epsilon}{0.6 + 4\epsilon}, \qquad \overline{P}_{\mathfrak{B}}(\mathrm{C} \mid Q_2) = \frac{0.5 + 5\epsilon}{0.6 + 4\epsilon},$$

$$\underline{P}_{\mathfrak{D}}(\mathrm{C} \mid Q_2) = \frac{0.2 - 2\epsilon}{0.7 + 3\epsilon}, \qquad \overline{P}_{\mathfrak{D}}(\mathrm{C} \mid Q_2) = \frac{0.5 + 5\epsilon}{0.7 + 3\epsilon},$$

$$\underline{P}_{\mathfrak{G}}(\mathrm{C} \mid Q_2) = \frac{1}{3}, \qquad \overline{P}_{\mathfrak{G}}(\mathrm{C} \mid Q_2) = \frac{2}{3}.$$

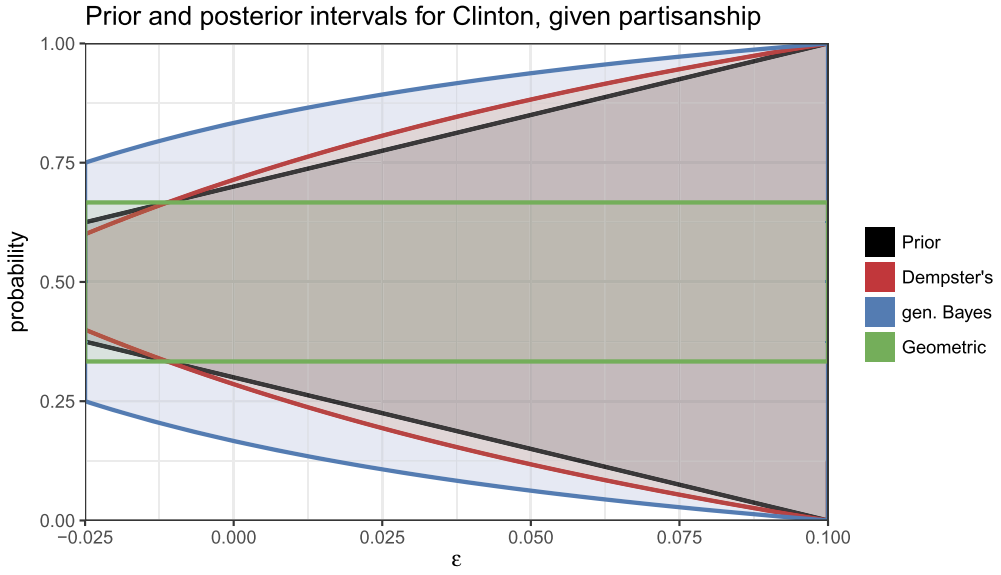

FIG. 2. *Prior probability interval for Clinton's voter support* (*black*) *and posterior probability intervals given reported partisanship according to the three updating rules* (*blue*: *generalized Bayes*, *red*: *Dempster's*, *green*: *Geometric*). *Due to full symmetry of the setup*, *contraction 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

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

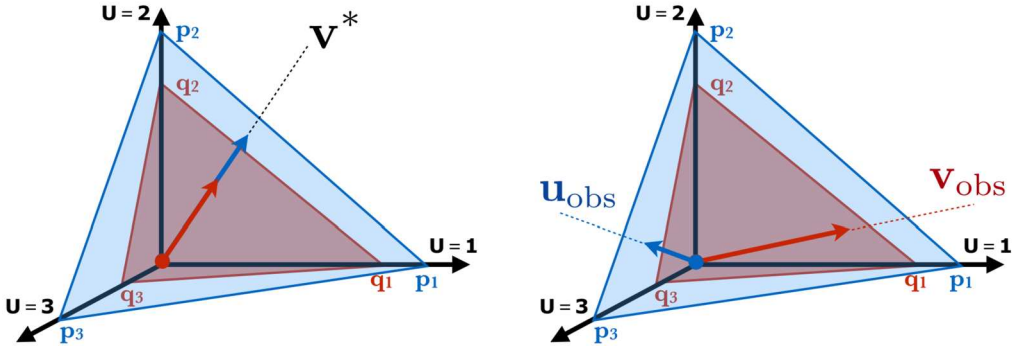

FIG. 3. *Ideal aggregating rules guarantee the comparison between treatment arms is made on a fair ground. Observed Simpson's paradox is strong evidence that the de facto aggregating rules are fair for comparison. Left*: *if* $p_k > q_k$ *for all* k, *then* $\mathbf{p}^\top \mathbf{v} > \mathbf{q}^\top \mathbf{v}$ *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 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. (2008). *Classic Works of the Dempster–Shafer Theory of Belief Functions* **219**. Springer, New York, NY.

THE ORIGINAL REFERENCE LIST

The list of entries below corresponds to the original Reference section of your article. The bibliography section on previous page was retrieved from MathSciNet applying an automated procedure. Please check both lists and indicate those entries which lead to mistaken sources in automatically generated Reference list.

Balch, M.S., Martin, R.and Ferson, S. 2017. Satellite conjunction analysis and the false confidence theorem. arXiv preprint arXiv:1706.08565.

Billingsley, P. 2013. Convergence of probability measures. John Wiley & Sons.

Blyth, C.R. 1972. On Simpson's paradox and the sure-thing principle. Journal of the American Statistical Association 67 338 364–366.

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. Journal of the National Cancer Institute 22 173–203.

Dempster, A.P. 1967. Upper and lower probabilities induced by a multivalued mapping. The Annals of Mathematical Statistics 38(2) 325–339.

Diaconis, P. 1978. Review of "A mathematical theory of evidence" (G. Shafer). Journal of the American Statistical Association 73(363) 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 Sensitivity analysis without assumptions. Epidemiology 27(3) 368.

Fagin, R.and Halpern, J.Y. 1991. A New Approach to Updating Beliefs A new approach to updating beliefs. In Proceedings of the Sixth Conference on Uncertainty in Artificial Intelligence Proceedings of the Sixth Conference on Uncertainty in Artificial Intelligence (317–325).

Fygenson, M. 2008. Modeling and predicting extrapolated probabilities with outlooks Modeling and predicting extrapolated probabilities with outlooks. Statistica Sinica 18(1) 9–40.

Gelman, A. 2006. The boxer, the wrestler, and the coin flip: a paradox of robust Bayesian inference and belief functions The boxer, the wrestler, and the coin flip: a paradox of robust Bayesian inference and belief functions. The American Statistician 60(2) 146–150.

Gong, R.and Meng, X.L. 2019. Probabilistic underpinning of belief functions for low-resolution statistical modeling. In preparation.

Good, I. 1974. A little learning can be dangerous A little learning can be dangerous. The British Journal for the Philosophy of Science 25(4) 340–342.

Hannig, J., Xie, M.g. 2012. A note on Dempster-Shafer recombination of confidence distributions A note on Dempster-Shafer recombination of confidence distributions. Electronic Journal of Statistics 6 1943–1966.

Hannig, J., Iyer, H., Lai, R.C.and Lee, T.C. 2016. Generalized fiducial inference: A review and new results Generalized fiducial inference: A review and new results. Journal of the American Statistical Association 111515 1346–1361.

Heitjan, D.F. 1994. Ignorability in general incomplete-data models Ignorability in general incomplete-data models. Biometrika 814 701–708.

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

Herron, T., Seidenfeld, T.and Wasserman, L. 1994. The extent of dilation of sets of probabilities and the asymptotics of robust Bayesian inference 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 (1994, 250–259).

Herron, T., Seidenfeld, T.and Wasserman, L. 1997. Divisive conditioning: further results on dilation Divisive conditioning: further results on dilation. Philosophy of Science 64(3) 411–444.

Huber, P.J.and Strassen, V. 1973. Minimax tests and the Neyman-Pearson lemma for capacities Minimax tests and the Neyman-Pearson lemma for capacities. The Annals of Statistics 1(2) 251–263.

Jaffray, J.Y. 1992. Bayesian updating and belief functions Bayesian updating and belief functions. IEEE transactions on Systems, Man, and Cybernetics 22(5) 1144–1152.

Kish, L. 1965. Survey sampling Survey sampling. John Wiley and Sons, New York, NY.

Kohlas, J. 1991. The reliability of reasoning with unreliable arguments The reliability of reasoning with unreliable arguments. Annals of Operations Research 32(1) 67–113.

Kruse, R., Schwecke, E. 1990. Specialization–a new concept for uncertainty handling with belief functions Specialization–a new concept for uncertainty handling with belief functions. International Journal Of General System18149–60.

Kyburg, H.E. 1987. Bayesian and non-Bayesian evidential updating Bayesian and non-Bayesian evidential updating. Artificial Intelligence313271–293.

Liu, K., Meng, X.L. 2014. Comment: A Fruitful Resolution to Simpson's Paradox via Multiresolution Inference Comment: A fruitful resolution to Simpson's Paradox via multiresolution inference. The American Statistician68117–29.

Liu, K., Meng, X.L. 2016. There Is Individualized Treatment. Why Not Individualized Inference? There is individualized treatment. Why not individualized inference? The Annual Review of Statistics and Its Applications379-111.

Martin, R., Liu, C. 2015. Inferential Models: reasoning with uncertainty Inferential Models: reasoning with uncertainty. CRC Press, Boca Raton, FL.

Meng, X.L., Xie, X. 2014. I got more data, my model is more refined, but my estimator is getting worse! Am I just dumb? I got more data, my model is more refined, but my estimator is getting worse! am i just dumb? Econometric Reviews331-4218–250.

Miranda, E., Montes, I. 2015. Coherent updating of non-additive measures Coherent updating of non-additive measures. International Journal of Approximate Reasoning56159–177.

Morgan, J.P., Chaganty, N.R., Dahiya, R.C.and Doviak, M.J. 1991. Let's make a deal: The player's dilemma Let's make a deal: The player's dilemma. The American Statistician454284–287.

Mosteller, F. 1965. Fifty challenging problems in probability with solutions Fifty challenging problems in probability with solutions. Courier Corporation, North Chelmsford, MA.

Nguyen, H.T. 1978. On random sets and belief functions On random sets and belief functions. Journal of Mathematical Analysis and Applications653531–542.

Pavlides, M.G., Perlman, M.D. 2009. How likely is Simpson's Paradox? How likely is Simpson's Paradox? The American Statistician633226–233.

Pearl, J. 1990. Reasoning with belief functions: An analysis of compatibility Reasoning with belief functions: An analysis of compatibility. International Journal of Approximate Reasoning45-6363–389.

Pedersen, A.P., Wheeler, G. 2014. Demystifying dilation Demystifying dilation. Erkenntnis7961305–1342.

Rubin, D.B. 1976. Inference and missing data Inference and missing data. Biometrika633581–592.

Schweder, T., Hjort, N.L. 2016. Confidence, Likelihood, Probability Confidence, likelihood, probability. Cambridge University Press, Cambridge, UK.

Seidenfeld, T., Wasserman, L. 1993. Dilation for sets of probabilities Dilation for sets of probabilities. The Annals of Statistics2131139–1154.

Shafer, G. 1976. A mathematical theory of evidence A mathematical theory of evidence. Princeton University Press, Princeton, NJ.

Shafer, G. 1979. Allocations of probability Allocations of probability. The Annals of Probability75827–839.

Shafer, G. 1985. Conditional probability Conditional probability. International Statistical Review/Revue Internationale de Statistique261–275.

Simpson, E.H. 1951. The interpretation of interaction in contingency tables The interpretation of interaction in contingency tables. Journal of the Royal Statistical Society. Series B (Methodological)132238–241.

Smets, P. 1991. About updating About updating. In Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence 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 Belief functions: The disjunctive rule of combination and the generalized Bayesian theorem. International Journal of Approximate Reasoning911–35.

Suppes, P., Zanotti, M. 1977. On using random relations to generate upper and lower probabilities On using random relations to generate upper and lower probabilities. Synthese364427–440.

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

Walley, P. 1991. Statistical reasoning with imprecise probabilities Statistical reasoning with imprecise probabilities. Taylor & Francis, Oxford, UK.

Wasserman, L., Kadane, J.B. 1990. Bayes theorem for Choquet capacities Bayes theorem for Choquet capacities. The Annals of Statistics1831328–1339.

Xie, M.G., Singh, K. 2013. Confidence distribution, the frequentist distribution estimator of a parameter: A review Confidence distribution, the frequentist distribution estimator of a parameter: A review. International Statistical Review8113–39.

Yager, R.R. 1987. On the Dempster-Shafer framework and new combination rules On the Dempster-Shafer framework and new combination rules. Information Sciences41293–137.

Yager, R.R., Liu, L. 2008. Classic works of the Dempster-Shafer theory of belief functions Classic works of the Dempster-Shafer theory of belief functions (219). Springer, New York, NY.

META DATA IN THE PDF FILE

Following information will be included as pdf file Document Properties:

Title : Judicious Judgment Meets Unsettling Updating: Dilation, Sure Loss and Simpson's Paradox
Author : Ruobin Gong, Xiao-Li Meng
Subject : Statistical Science, 0, Vol. 0, No. 00, 1-22
Keywords: Imprecise probability, model uncertainty, Choquet capacity, belief function, coherence, Monty Hall problem
Affiliation:

THE LIST OF URI ADDRESSES

Listed below are all uri addresses found in your paper. The non-active uri addresses, if any, are indicated as ERROR. Please check and update the list where necessary. The e-mail addresses are not checked – they are listed just for your information. More information can be found in the support page:
http://www.e-publications.org/ims/support/urihelp.html.

301 http://www.imstat.org/sts/ [2:pp.1,1] Moved Permanently
301 http://www.imstat.org [2:pp.1,1] Moved Permanently // http://www.imstat.org/
--- mailto:rg915@stat.rutgers.edu [2:pp.1,1] Check skip
--- mailto:meng@stat.harvard.edu [2:pp.1,1] Check skip

On the history and limitations of probability updating

Glenn Shafer

Abstract. Gong and Meng show that we can gain insights into classical paradoxes about conditional probability by acknowledging that apparently precise probabilities live within a larger world of imprecise probability. They also show that the notion of updating becomes problematic in this larger world. A closer look at the historical development of the notion of updating can give us further insights into its limitations.

Key words and phrases: Bayes's rule of conditioning, Dempster's rule, conditional probability, conditionalization, imprecise probabilities, probability protocols, relative probability, updating.

1. A BROADER PERSPECTIVE ON CLASSICAL PARADOXES

Conditional probability paradoxes, stories in which $P(A|B)$ does not seem to be a reasonable probability for A after we learn B, have been with us since the late 19th century.[1] Many of these paradoxes turn on initial probabilities not telling us enough about the relation between A and the event that we learn B. Many authors have explained this, but each in their own way, often vociferously denying the cogency of others' explanations. No consensus having emerged, the paradoxes endure.

Roubin Gong and Xiao-Li Meng propose a broader perspective. Instead of trying to resolve the paradoxes within standard probability theory, in which we have joint probabilities for all events of interest, they propose that we use the theory of imprecise probabilities, in which events of interest may have only upper and lower probabilities and quantities of interest may have only upper and lower expected values. The theory of imprecise probabilities not only generalizes the standard theory but also allows us to recognize formally the incompleteness of any standard (a.k.a. "precise") probability model. We do this by adding events to the model without adding probabilities for them, thus obtaining a larger "imprecise" model. As Gong and Meng put it,

> 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.

This allows them to explain the conditional probability paradoxes this way:

> Their narratives imply the existence of a joint distribution, yet only a subset of marginal information is precisely specified.

The theory of imprecise probability has flourished for several decades, but largely outside statistics journals. Bringing it into the statistical mainstream, as Gong and Meng have done with this article in *Statistical Science*, is a welcome move. As Gong and Meng show, the theory's ideas can enrich statisticians' understanding of longstanding questions within our community. We can also hope that the critical resources of the statistical community can add new depth to the theory. Gong and Meng tell us that dilation, contraction, and sure loss "hint at novel types of information contribution". Perhaps we need theories of these novel types.

University Professor, Rutgers University, Newark, New Jersey, (e-mail: gshafer@rutgers.edu).

[1] See Bertrand (1889). Bertrand's paradoxes have been discussed by Shafer and Vovk (2003), Gorroochurn (2012), and many others.

2. DO ALL EVENTS HAVE NUMERICAL PROBABILITIES?

The theory of imprecise probabilities says no. Many events, perhaps most, do not have numerical probabilities. Is this a new or controversial view?

Certainly it is not new. Before the 18th century, scholars who wrote about degrees of probability seldom suggested that these degrees could ever be put in numerical form (Knebel, 2000). Before 1713, when Jacob Bernoulli's *Ars conjectandi* appeared, even expectations in games of chance were not usually connected with the idea of probability. Bernoulli made the connection and launched the project of finding numerical probabilities not only for games of chance but also for civil, criminal and business matters. But Bernoulli did not believe that we can always find probabilities for a thing and its contrary that add to one.

Jean Le Rond d'Alembert, the uncontested leader of French mathematics in his time, was an avowed skeptic about Bernoulli's ambition for numerical probabilities. In 1676, the same year the teenage Pierre-Simon Laplace arrived in Paris seeking his patronage, d'Alembert published his own views about the art of conjecture. According to d'Alembert, this art has three branches (D'Alembert, 1767, Chapter VI):

1. The first branch is games of chance. Here we can count equally likely cases and reason about them *a priori*.[2]
2. The second consists of topics such as insurance and inoculation, where we can learn the number of cases and their ratios only from experience and only approximately.
3. The third consists of the many topics for which mathematical demonstration is rare or impossible. D'Alembert included here physics, history, medicine, the law, and business.

Outside the small world of scholars who specialize in mathematical probability and its applications, these views probably found widespread assent when d'Alembert published them and may continue to do so today. Over time, scientists and statisticians may have moved bits of d'Alembert's third category into the second or even the first, but the third still seems very large.

When I began my own study of mathematical statistics in the early 1970s, I took it for granted that only some events have probabilities. Both R. A. Fisher and Andrei Kolmogorov had said so explicitly.[3] I thought nearly all statisticians, philosophers, and mathematicians agreed. Today I am not so sure. For decades now, Bayesians have insisted that a person can supply a personal probability for anything. As realism has gained ascendancy in philosophy, the claim that anything uncertain has an objective probability, usually unknown, has also become common. Many physicists now imagine a universal wave function. Many mathematical probabilists now imagine the whole course of the world being described by a single element ω of a vast probability space Ω. In this context, I am tempted to see the increasing popularity of the theory of imprecise probabilities as a return to d'Alembert's common sense.

3. MODEL OR JUDGMENT?

As leaders in the "Bayesian, Fiducial & Frequentist" community, Gong and Meng want to transcend the quarrels between proponents of different interpretations of probability and different methodologies for statistical inference. This is visible in their choice of words. They avoid saying whether the probabilities they discuss, precise or imprecise, are objective facts or subjective beliefs, and they make heavy use of the word "model". The first two sentences of their article reveal, however, that the models being studied are akin to neo-Bayesians. They update themselves:

---

[2] D'Alembert was also skeptical about some of this *a priori* reasoning. Can you really know *a priori* the probability of getting a head tossing a coin when you are allowed to try twice? As Bernard Bru has argued, we should hesitate to dismiss d'Alembert's doubts on this point as a mere "gambler's fallacy" (Bru, 1989, 2002).

[3] Kolmogorov's most explicit statement that not every event has a probability may be in his article on probability in the 1951 edition of the *Great Soviet Encyclopedia* (Shafer and Vovk, 2003, p. 50). Fisher was equally explicit, stating in 1956, for example, that "in some cases no probability exists" (Fisher, 1956, p. 45).

> Statistical learning is a process through which models perform updates in light of new information, according to a pre-specified 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.

By the end of the article, however, I was wondering whether these first two sentences were a declaration of faith or a straw man. Is "judicious judgment" limited to choosing an updating rule before the fact, incorporating it into the model, and letting the model do our later thinking for us? Or is "judicious judgment" most needed after something unexpected is observed? I would welcome the second interpretation and see it as another step back to common sense.

4. FROM RELATIVE TO CONDITIONAL PROBABILITY

Two centuries before the formula

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)} \tag{1}$$

became a definition, Abraham De Moivre provided a betting argument for what became known as "the rule of compound probability": the probability of two events both happening is the probability of the first times the probability of the second "when the first shall have been consider'd as having happen'd" (De Moivre, 1738, p. 7). As this formulation reveals, De Moivre did not begin with a probability measure that gave joint probabilities for all events he wanted to discuss. Instead he constructed joint probabilities from simpler ingredients. The probability of a second event given the first was one of these ingredients. It was not a "conditional probability"; it was the probability of the second event in the new situation in a betting game. The rule of compound probability remained one of the basic rules of probability theory until the mid-20th century, when mathematical probabilists decided that it was more convenient to make probability measures their starting point, thus shifting (1) from being a consequence of the rule of compound probability to being a definition of $P(\cdot|\cdot)$.

Nineteenth-century mathematicians sometimes wrote about "relative probability". In his popular French textbook on probability, first published in 1816, Sylvestre-François Lacroix called the ratio $P(A)/(P(A)+P(B))$ the *probabilité relative* of A as compared to B . When rolling two dice for example, where there are 6 chances for getting a 4 and only 3 chances for getting a 4, the probability of 7 relative to 4 is $2/3$ (Lacroix, 1816, pp.19–20). We see this same notion of relative probability in (Liagre, 1852, §16).

It seems that "conditional probability" first appeared in George Boole's *Laws of Thought* (Boole, 1854, Ch. XX, §21). A logician, Boole was trying to make mathematical probability part of logic, and he was accustomed to using "condition" and "conditional" in logic. Boole's used "conditional probability" only once, however, casually and perhaps even inadvertently, as he was writing mostly about "conditional events". In 1887, in his Metretike, Francis Edgeworth, citing Boole, systematically called the probability of an effect given a cause a "conditional probability" (Mirowski, 1994). We already see the German and Russian equivalents, *bedingte Wahrscheinlichkeit* and условная вероятность, in the early 20th century (Shafer and Vovk, 2003, p. 6).

In the course of commenting on Boole, Charles Sanders Peirce wrote, "Let b_{a} denote the frequency of b's among the a's" (Peirce, 1867, p. 255). Because Peirce was identifying probability with frequency, this could be considered the first notation for conditional probability. Others made other suggestions, mostly independently of each other. Hugh McColl, independently of Peirce, wrote "The symbol x_a denotes the chance that the statement a is true on the assumption that the statement a is true" (McColl, 1880, 1881). Later he used $\frac{A}{B}$ (MacColl, 1897). Andrei Markov (1900) wrote (A, B).

In 1911, John Maynard Keynes introduced what he called "the fundamental symbol of probability", A/H, for the probability of A given H. This symbol became popular at Cambridge; we see it in books by C. D. Broad (Broad, 1914, p. 318), John Maynard Keynes (Keynes, 1921, p. 177), and William E. Johnson (Johnson, 1924, p. 179). To all appearances,

Keynes first used the symbol in the 1908 dissertation that grew into his book, and Johnson popularized it in conversations and lectures.[4]

In 1901, the German mathematician Felix Hausdorff introduced the symbol $P_F(E)$ for what he called the *relative Wahrsheinlichkeit von E, posito F* (relative probability of E given F). In his view, the absolute probability $P(E)$ of an event E is simply the relative probability $P_F(E)$, where F is our current knowledge. This knowledge can change, and Hausdorff mentioned three examples (Hausdorff, 1901, pp. 154–155):

- When the absolute probability $P(E)$ is a weighted average of possible objective probabilities, F represents one of the possible objective probabilities, and we learn that F is correct, then we change $P(E)$ to $P_F(E)$.
- We may learn that there were more possibilities than we had realized, as when we learn that the geometry of the world may not be Euclidean. In this case, we shift from $P_F(E)$ to $P_G(E)$, where G permits this wider set of possibilities.
- We may learn that our knowledge F was flatly wrong and therefore shift from $P_F(E)$ to $P_G(E)$, where G contradicts F.

Emmanuel Czuber followed Hausdorff's terminology and notation in the second edition of his influential textbook, except that he used $\mathfrak{W}_F(E)$ instead of $P_F(E)$ (Czuber, 1908, pp. 44–45). Kolmogorov used $P_A(B)$ in his pathbreaking 1933 *Grundbegriffe*, but he called such a probability *bedingte* (conditional), not *relative* as Hausdorff and Czuber had done (Kolmogorov, 1933, p. 206).

Our current notation $P(\cdot|\cdot)$ is apparently due to Harold Jeffreys. In 1919, Dorothy Wrinch and Jeffreys had used $P(p : q)$ (Wrinch and Jeffreys, 1919). In 1931, Jeffreys replaced this with $P(p|q)$, commenting on its advantage over $P(p : q)$ and notation p/q in a way that makes clear that he was not aware of any previous use of $P(p|q)$ (Jeffreys, 1931, p. 31).

5. FROM CONDITIONAL PROBABILITY TO UPDATING

After World War II, mathematicians, statisticians, and philosophers began to take it for granted that the proper setting for mathematical probability is a probability measure rather than a collection of probabilities less structured or structured in some other way. Only then did it become natural to recast the notion of conditional probability as an action with probabilities as its object: a statistician or scientist "conditionalizes" or "conditions" or "updates" the probabilities. This formulation seems to have slipped unheralded into many minds. The earliest instance of it I have found is in Estes and Suppes (1957). After emphasizing the importance for psychology of the concept of a probability measure (p. 11), Estes and Suppes explained that "the experimenter may conditionalize the probabilities of reinforcement upon preceding events of the sample space in whatever manner he pleases" (pp. 20–21). The use of "update" in this context seems to have appeared much later, only in the late 1970s.

In the 1960s, A. P. Dempster was writing about his own rules for or of combination and conditioning and comparing them with Bayesian rules (Dempster, 1967, 1968). In my 1976 book on the Dempster-Shafer theory (Shafer, 1976), I distinguished between *Bayes's rule of conditioning*, as I called it, and Bayes's theorem.

- Bayes's rule of conditioning says that when you learn A, you change your probability for B from $P(B)$ to $P(B|A)$ as given by (1), regardless of the order in which the events may have happened in the world. I attributed this rule directly to Bayes because he had given a betting argument for it, which is erroneous in my opinion; see Shafer (1982).

[4] Keynes claimed originality for the symbol in correspondence with W. H. Macaulay in 1907 (Aldrich, 2020). In his book he says that had not been aware of McColl's earlier notation when he devised the symbol (Keynes, 1921, p. 177). In a review of Keynes's book, Broad suggested that Keynes had borrowed the symbol from Johnson (Broad, 1922, p. 78), but Johnson acknowledged Keynes's priority, at least in publication. Johnson read Keynes's dissertation and likely used Keynes's symbol subsequently in lectures attended by Broad and Dorothy Wrinch (Aldrich, 2008, 2020).

- Bayes's theorem is more specific; it is the Bayesian rule for changing probabilities for a parameter based on observations (or, in Laplace's words, obtaining probabilities for causes from events). Beginning with Cournot (1843), some authors called this Bayes's rule (*règle de Bayes* in French; *Bayesschen Regel* in German); others called it Bayes's formula or Bayes's theorem.[5] In English, it was often called the method of "inverse probability". Now that (1) is regarded as a definition, it is more often called a theorem.

The distinction between Bayes's rule of conditioning (or updating or conditionalization; see Teller (1973)) and Bayes's theorem is now widely made, but it is remains unfamiliar to many statisticians. Perhaps for this reason, Gong and Meng blur the distinction, calling

$$P(A|B) = P(A)\frac{P(B|A)}{P(B)}$$

"Bayes rule". I find this confusing, because when (1) is treated as a general rule for updating a probability measure after observing an event, there is no presumption that the probabilities of the event conditional on all other events had previously been singled out and calculated.

6. THE IMPLICATIONS OF INSISTING ON A PROTOCOL

Gong and Meng are kind enough to cite the 1985 article in which I insisted that Bayesian updating after learning B is legitimate only in the presence of a protocol that singled out B as one of the things we might learn (Shafer, 1985). It is only in this case, I argued, that De Moivre's betting argument and its variants (e.g., de Finetti (1937); Teller (1973)) justify Bayes's rule of conditioning and only in this case that paradox can be avoided. I would like to add to their discussion an explanation of how I understand the consequences of insisting on a protocol.

By a *protocol*, I mean what Joseph L. Doob and later probabilists have called a *filtration*. Starting at time 0, you first learn X_1, then X_2, etc. In the simple special case where these variables are all binary and we stop at fixed time n, we can visualize the protocol as a binary tree. The sample space Ω is the set of all paths through the tree, from time 0 to time n. There are 2^n elements in Ω and hence 2^{2^n} events. But there are exponentially fewer nodes in the tree — only $2^n - 1$. But only a node in the tree can represent what you may have learned at some point in time. If and when you reach a particular node, say by observing $x_1, \ldots, x_k$, your new probability for an event A will be your original probability "conditioned" on $X_1 = x_1, \ldots, X_k = x_k$. But you will never "condition" on any of the $2^{2^n} - 2^n + 1$ events not of this form. So the notion that you have a methodology that allows you to "update" when your new information is any subset B of Ω is illusory.

A common Bayesian response is that you should of course condition on everything you have learned, including the fact that you learned it. This implies that the elements of Ω specify what you will and will not learn at every point in time. So the Bayesian view already implicitly calls for a protocol for how new information may arrive. In my view, leaving this need for a protocol implicit is more than an invitation to paradox. It is deceptive. Once the demand to provide a probability model for your entire learning process is made explicit, it becomes obvious that the demand often cannot be satisfied.

Surely we should conclude that models with updating rules are only one limited set of tools for assessing uncertainty. We also need ideas for evaluating and combining unanticipated evidence, such as Jacob Bernoulli proposed in (Bernoulli, 1713, 2006, Part IV, Ch. 3), Dempster and I proposed in the 1960s and 1970s, and others have proposed before and since.

7. ACKNOWLEDGEMENTS

My preparation of this note has benefited from recent conversations with John Aldrich, Bernard Bru, Roubin Gong, Xiao-Li Meng, and Sandy Zabell, and from countless conversations over the years with other colleagues.

[5] Bayes's friend and executor Richard Price used the phrase "Mr. Bayes's rules" to refer to formulas Bayes had derived for approximating what we now call posterior and predictive Bayesian probabilities in the binomial problem (Dale, 1999, pp. 39–40).

REFERENCES

ALDRICH, J. (2008). Keynes among the statisticians. *History of Political Economy* **40** 265–316.

ALDRICH, J. (2020). Personal communication.

BERNOULLI, J. (1713). *Ars Conjectandi*. Thurnisius, Basel. See (Bernoulli, 2006, translation by Edith Sylla).

BERNOULLI, J. (2006). *The Art of Conjecturing, together with Letter to a Friend on Sets in Court Tennis*. Johns Hopkins University Press, Baltimore.

BERTRAND, J. (1889). *Calcul des probabilités*. Gauthier-Villars, Paris.

BOOLE, G. (1854). *An investigation of the laws of thought, on which are founded the mathematical theories of logic and probabilities*. Macmillan, London. Reprinted by Dover, New York, 1958.

BROAD, C. D. (1914). *Perception, Physics and Reality: An Enquiry into the Information that Physical Science can Supply about the Real*. Cambridge University Press.

BROAD, C. D. (1922). *A Treatise on Probability*. By J. M. Keynes. *Mind* **31** 72–85.

BRU, B. (1989). Doutes de d'Alembert sur le calcul des probabilités. In *Jean d'Alembert, savant et philosophe. Portrait à plusieurs voix* (M. Emery and P. Monzani, eds.) 279–292. Archives contemporaines, Paris.

BRU, B. (2002). Des fraises et des oranges. In *Sciences, musiques, lumières: Mélanges offerts à Anne-Marie Chouillet* (U. Kölving and I. Passeron, eds.) 3–10. Centre international d'étude du XVIII^{e} siècle, Ferny-Voltaire.

COURNOT, A. A. (1843). *Exposition de la théorie des chances et des probabilités*. Hachette, Paris. Reprinted in 1984 as Volume I (Bernard Bru, editor) of Cournot (1973–2010).

COURNOT, A. A. (1973–2010). *Œuvres complètes*. Vrin, Paris. The volumes are numbered I through XI, but VI and XI are double volumes.

CZUBER, E. (1908). *Wahrscheinlichkeitsrechnung und ihre Anwendung auf Fehlerausgleichung, Statistik und Lebensversicherung* **1**, 2nd ed. Teubner, Leipzig.

DALE, A. I. (1999). *A History of Inverse Probability From Thomas Bayes to Karl Pearson*, Second ed. Springer, New York.

D'ALEMBERT, J. L. R. (1767). *Éclaircissmens sur différens endroits des* Élémens *de philosophie*. Chatelain, Amsterdam. In the 5th volume of d'Alembert's *Mélanges de littéraire d'histoire et de philosophie*.

DE FINETTI, B. (1937). La prévision: ses lois logiques, ses sources subjectives. *Annales de l'institut Henri Poincaré* **7** 1–68.

DE MOIVRE, A. (1738). *The Doctrine of Chances: or, A Method of Calculating the Probabilities of Events in Play*, 2nd ed. Pearson, London.

DEMPSTER, A. P. (1967). Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics* **38** 325–339.

DEMPSTER, A. P. (1968). A generalization of Bayesian inference (with discussion). *Journal of the Royal Statistical Society B* **30** 205–247.

ESTES, W. K. and SUPPES, P. (1957). Foundations of Statistical Learning Theory, I. The linear model for simple learning. Technical Report No. 16, Behavioral Sciences Division, Applied Mathematics and Statistics Laboratory, Stanford University.

FISHER, R. A. (1956). *Statistical Methods and Scientific Inference*. Oliver and Boyd, Edinburgh. Subsequent editions appeared in 1959 and 1973.

GORROOCHURN, P. (2012). *Classic Problems of Probability*. Wiley, New York.

HAUSDORFF, F. (1901). Beiträge zur Wahrscheinlichkeitsrechnung. *Sitzungsberichte der Königlich Sächsischen Gesellschaft der Wissenschaften zu Leipzig, Mathematisch-Physische Klasse* **53** 152–178.

JEFFREYS, H. (1931). *Scientific Inference*, 1st ed. Cambridge University Press.

JOHNSON, W. E. (1924). *Logic* **3**. Cambridge University Press.

KEYNES, J. M. (1921). *A Treatise on Probability*. Macmillan, London.

KNEBEL, S. K. (2000). *Wille, Würfel und Wahrscheinlichkeit: Das System der moralischen Notwendigkeit in der Jesuitenscholastik 1550–1700*. Meiner, Hamburg.

KOLMOGOROV, A. N. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer, Berlin. English translation *Foundations of the Theory of Probability*, Chelsea, New York, 1950, 2nd ed. 1956.

LACROIX, S. F. (1816). *Traité élémentaire du calcul des probabilités*. Courcier, Paris. Second edition 1822.

LIAGRE, J.-B.-J. (1852). *Calcul des probabilités et théorie des erreurs avec des applications aux sciences d'observation en général et à la géodésie*. Muquardt, Brussels. Second edition, 1879, prepared with the assistance of Camille Peny.

MACCOLL, H. (1897). The calculus of equivalent statements (sixth paper). *Proceedings of the London Mathematical Society* **28** 555–579.

MARKOV, A. A. (1900). *Probability Calculus (in Russian)*. Imperial Academy, St. Petersburg. The second edition, which appeared in 1908, was translated into German as *Wahrscheinlichkeitsrechnung*, Teubner, Leipzig, Germany, 1912.

MCCOLL, H. (1880). The calculus of equivalent statements (fourth paper). *Proceedings of the London Mathematical Society* **11** 112–121.

MCCOLL, H. (1881). A note on Prof. C. S. Peirce's probability notation of 1867. *Proceedings of the London Mathematical Society* **12** 102.

MIROWSKI, P., ed. (1994). *Edgeworth on Chance, Economic Hazard, and Statistics*. Rowman & Littlefield, Lanham, Maryland.

PEIRCE, C. S. (1867). On an improvement in Boole's calculus of logic. *Proceedings of the American Academy of Arts and Sciences* **7** 250–261.

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

SHAFER, G. (1982). Bayes's two arguments for the rule of conditioning. *Annals of Statistics* **10** 1075–1089.

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

SHAFER, G. and VOVK, V. (2003). The origins and legacy of Kolmogorov's *Grundbegriffe*. Working Paper 4, www.probabilityandfinance.com.

TELLER, P. (1973). Conditionalization and observation. *Synthese* **26** 218–258.

WRINCH, D. and JEFFREYS, H. (1919). On some aspects of the Theory of Probability. *Philosophical Magazine* **38** 715–731.

Submitted to Statistical Science

Comment on Gong and Meng's "Judicious judgment meets unsettling updating: Dilation, sure loss, and Simpson's paradox"

Chuanhai Liu and Ryan Martin

Abstract. Here we demonstrate that the *inferential model* (IM) framework, unlike the updating rules that Gong and Meng show to be unreliable, provides valid and efficient inferences/prediction while not being susceptible to sure loss. In this sense, the IM framework settles what Gong and Meng characterized as "unsettling."

Key words and phrases: belief function, efficiency, lower and upper probability, inferential models, validity.

1. INTRODUCTION

Ruobin Gong and Xiao-Li Meng are to be congratulated for their thought-provoking article shedding light on the paradoxical results that can surface when imprecise or incompletely-specified models are updated, in light of observed data, using formal rules like Dempster's and generalized Bayes. With scientific problems becoming increasingly more complex, the idea that models describing the phenomena under investigation can be precisely specified is a fantasy, so Gong and Meng's insights about the effects of these updating rules are both important and timely. However, after highlighting a number of cases where the updates are "unsettling," they give no recommendations about which updating rule, if any, is reliable. In some cases, generalized Bayes seems to be the right choice, while in others it's Dempster's rule. Since we can't rely on any of the updating rules to give satisfactory answers in every problem, apparently our only recourse is to use "judicious judgment" on a case-by-case basis.

Here we argue that steps toward settling what's unsettling about these updates can be made by taking a different perspective on what a solution to the problem entails. Gong and Meng make their perspective very clear:

> *Statistical learning is a process through which models perform updates in light of new information, according to a pre-specified set of operation rules.*

Chuanhai Liu is Professor, Department of Statistics, Purdue University, West Lafayette, Indiana, USA (e-mail: chuanhai@purdue.edu). Ryan Martin is Professor, Department of Statistics, North Carolina State University, Raleigh, North Carolina, USA (e-mail: rgmarti3@ncsu.edu).

What's missing from this description is that inferences drawn based on the updated models must be reliable or valid in some specific sense, otherwise, the results are not useful. So the question is not really about updating beliefs but, rather, how to ensure that the beliefs data scientists construct for inference and prediction achieve the desired reliability properties. From this perspective, Gong and Meng's goal is overly ambitious: for valid and efficient inference, rules that update beliefs are not necessary. A less ambitious goal—but still in line with the priorities of scientists—is to understand what it takes to construct procedures for allocating beliefs such that inferences drawn are *valid* and *efficient*. The first step is to define what these terms mean, which we do below in Section 2. We immediately take comfort in the fact that validity rules out the troubling sure loss phenomenon, and, as we show in Section 3, validity and efficiency make it possible to compare the solutions based on the different updating rules. Of course, if validity and efficiency are the goal, then it makes sense to follow a procedure that is specifically design to achieve these properties. The *inferential model* (IM) procedure introduced in Martin and Liu (2013, 2015a) is just that, and in Section 4 we describe this framework and show how it generally leads to better solutions than those based on the formal updating rules in Gong and Meng's examples. The take-away message is that, by following the validity- and efficiency-focused IM approach, the "unsettling" phenomena identified by Gong and Meng can be avoided. Finally, Section 5 concludes with few topics for future investigation.

2. VALID AND EFFICIENT PREDICTION

The examples in Gong and Meng (2020) are most conveniently described as prediction problems, so that's the

perspective we take; all of this can be developed in a similar way for inference. To set the scene, let X denote the observable data and $Y \in \mathbb{Y}$ the quantity to be predicted. Next, let P denote the probability measure that describes the joint distribution of (X, Y), at least partially unknown or unspecified. As indicated above, we proceed by quantifying uncertainty about Y, given $X = x$, via a pair of lower and upper probabilities, denoted by $(\underline{\pi}_x, \overline{\pi}_x)$, defined on $\mathbb{Y}$. We refer to the map $x \mapsto (\underline{\pi}_x, \overline{\pi}_x)$ as a *probabilistic predictor*, and the user's degree of belief in the truthfulness of an assertion $A \subseteq \mathbb{Y}$ concerning the unobserved Y, given $X = x$, are described by the pair $(\underline{\pi}_x(A), \overline{\pi}_x(A))$. Note that the probabilistic predictor need not be based on updating a precise or imprecise probability model.

Since the goal is for the probabilistic predictor to make reliable predictions, i.e., not wrong too often, consider the following prediction validity property.

DEFINITION (Cella and Martin 2020). A probabilistic predictor is *valid* if

$$\mathsf{P}\{\overline{\pi}_X(A) \leq \alpha, Y \in A\} \leq \alpha, \quad \forall\, (A, \alpha, \mathsf{P}), \tag{1}$$

where the probability is with respect to the joint distribution of (X, Y) determined by P and "$\forall$" is over all assertions $A \subseteq \mathbb{Y}$, all levels $\alpha \in [0, 1]$, and all P.

The intuition is that, at least for small α, the data analyst interprets the event "$\overline{\pi}_X(A) \leq \alpha$" as evidence against the truthfulness of the assertion A about Y, so the joint event "$\overline{\pi}_X(A) \leq \alpha, Y \in A$" is one where an erroneous prediction is possible. Then (1) requires that the user be able to control the frequency of such erroneous predictions. Thanks to the familiar duality between lower and upper probabilities, a similar condition can be formulated in terms of $\underline{\pi}_x$ (Cella and Martin, 2020). To see what condition (1) imposes on the probabilistic predictor, consider the equivalent expression

$$\mathsf{E}\{1_{\overline{\pi}_X(A) \leq \alpha}\, \mathsf{P}(Y \in A \mid X)\} \leq \alpha, \quad \forall\, (A, \alpha, \mathsf{P}), \tag{2}$$

where 1_B is the indicator function, E is expectation with respect to the marginal distribution of X under P, and $\mathsf{P}(Y \in A \mid X)$ is the conditional probability based on P. Clearly, if $\overline{\pi}_x(A)$ equals or dominates the conditional probability $\mathsf{P}(Y \in A \mid x)$ or the marginal probability $\mathsf{P}(Y \in A)$, then (2) holds. This connection between validity and "dominance" leads to several interesting observations, as discussed in Cella and Martin (2020).

- Sure loss, the most unsettling of the three phenomena studied by Gong and Meng, is ruled out by validity, that is, validity implies no sure loss.
- If the imprecise model is known to contain the true joint distribution of (X, Y), like in Gong and Meng's examples, then the generalized Bayes solution is valid.

While generalized Bayes provides a strategy to achieve validity, it's not the only option and often will not be the best; see below.

Beyond validity, efficiency is important too. Here, we say that between a pair of valid probabilistic predictors, with upper probabilities $\overline{\pi}_x$ and $\overline{\pi}'_x$, the latter is no less efficient than the former—with respect to a specified assertion A—if $\overline{\pi}'_x(A) \leq \overline{\pi}_x(A)$ for all x. The idea is that large upper probabilities are trivially valid, so the goal is to find the smallest possible upper probabilities that satisfy (1) or (2). By the duality between lower and upper probabilities, similar intuition can be developed for $\underline{\pi}_x$. We'll not investigate validity or efficiency formally here, only in the context of two examples in Section 3.

3. GONG AND MENG'S EXAMPLES

3.1 Three prisoners

Three prisoners—labeled A, B, and C—are in custody and one will be randomly chosen to have their sentence pardoned; the other two will be executed. Let Y denote the pardoned prisoner. Prisoner A ask the guard to tell him which of Prisoners B or C will be executed, and the guard's response is the data X. The goal is to predict Y based on data X. What do validity and efficiency add to the discussion?

As Gong and Meng argue, the joint distribution for (X, Y) is fully determined except for the conditional probability $\theta = \mathsf{P}(X = \text{B} \mid Y = \text{A})$. So, for the most relevant assertion, "$Y = \text{A}$," the validity condition (2) can be expressed as

$$1_{\overline{\pi}_{\text{B}}(\text{A}) \leq \alpha} \cdot \tfrac{\theta}{3} + 1_{\overline{\pi}_{\text{C}}(\text{A}) \leq \alpha} \cdot \tfrac{1-\theta}{3} \leq \alpha. \tag{3}$$

As presented in Gong and Meng—see, also, Walley (1991, Sec. 6.4.4)—the generalized Bayes solution returns a probabilistic predictor with

$$\underline{\pi}_x(\text{A}) = 0 \quad \text{and} \quad \overline{\pi}_x(\text{A}) = \tfrac{1}{2}, \quad x \in \{\text{B}, \text{C}\},$$

and, for this, it's easy to check that (3) holds. Dempster's rule returns a probabilistic predictor with *lower and upper* probabilities for "$Y = \text{A}$" equal to $\frac{1}{2}$, for all x. This satisfies (3) at "$Y = \text{A}$," but not if we consider the complementary assertion. Indeed, with Dempster's probabilistic predictor at the assertion "$Y \in \{\text{B},\text{C}\}$,", the validity requirement in (3) boils down to

$$1_{\frac{1}{2} \leq \alpha} \cdot \tfrac{2}{3} \leq \alpha. \tag{4}$$

Taking $\alpha = \frac{1}{2}$ leads to a contradiction. This is basically the proof of how sure loss leads to a violation of validity in general. Similarly, the solution based on the geometric rule, which also suffers from sure loss in this example, is invalid.

A closer look at (3) provides some insight as to what the "most efficient" solution is. If $\underline{\pi}_x(\text{A}) = \frac{1}{3}$ for each

$x \in \{\text{B}, \text{C}\}$, then (3) would be satisfied, and it would be more efficient than the generalized Bayes solution. It would also be valid since lower probability on the complementary event is $\frac{2}{3}$, as opposed to Dempster's $\frac{1}{2}$, so it would not get caught by the trap (4). We'll see below how this "most efficient" solution can be achieved.

3.2 Boxer, wrestler, and coin

Let Y_1 denote the outcome a fair coin flip, with $Y_1 = 1$ and $Y_1 = 0$ corresponding to Heads and Tails, respectively, and let Y_2 denote the outcome of the boxer versus wrestler match, with $Y_1 = 1$ and $Y_1 = 0$ denoting a boxer and wrestler victory, respectively. The data is $X = |Y_1 - Y_2|$, an indicator that Y_1 and Y_2 take the same value. The goal is to predict the outcome of the fight (or of the coin flip) based on the observed value of X.

Features of the joint distribution of (X, Y), with $Y = (Y_1, Y_2)$, are left unspecified, in particular, the conditional probabilities

$$\theta_{1|y_1} = \mathsf{P}(Y_2 = 1 \mid Y_1 = y_1), \quad y_1 \in \{0, 1\}.$$

This pair $\theta = (\theta_{1|0}, \theta_{1|1})$ of conditional probabilities can take any value in $[0,1]^2$. That is, the problem setup doesn't rule out the possibility that the fight's outcome is determined by the coin flip, or that the fight's outcome is independent of the coin and pre-determined.

As above, let's start by specializing the validity condition to the present example. That is, if $\overline{\pi}_x(1)$ is the probabilistic predictor's upper probability at the assertion "$Y_2 = 1$," i.e., a boxer victory, then (2) requires

$$\tfrac{1}{2}\big\{1_{\overline{\pi}_0(1) \le \alpha} \cdot \theta_{1|0} + 1_{\overline{\pi}_1(1) \le \alpha} \cdot \theta_{1|1}\big\} \le \alpha.$$

Since $(\theta_{1|0}, \theta_{1|1})$ can take any value in $[0,1]^2$, there is no way to ensure that validity holds, except trivially, by taking the upper probabilities identically equal to 1. This is precisely the generalized Bayes solution in Gong and Meng. Dempster's rule, again, is invalid.

For assertions about the coin, the only satisfactory solution based on the methods investigated in Gong and Meng is that based on Dempster's rule, which ignores the data and uses the known marginal distribution of Y_1. It's easy to check that the simple probabilistic predictor

$$\underline{\pi}_x(\text{"}Y_1 = 1\text{"}) = \overline{\pi}_x(\text{"}Y_1 = 1\text{"}) = \tfrac{1}{2}, \quad x \in \{0, 1\},$$

is valid and efficient. We'll see below how this solution can be achieved in the IM context.

4. INFERENTIAL MODELS

4.1 Formulation

The IM formulation starts by specifying an *association* between what is being modeled, i.e., data X and quantity of interest Y, the unknown parameter $\theta \in \Theta$, and an unobservable auxiliary variable U, whose distribution P_U is known, via an equation or rule

$$(X, Y) = a(\theta, U), \quad U \sim \mathsf{P}_U. \tag{5}$$

The mapping $a(\theta, \cdot)$ implicitly encodes what is known about the joint distribution but explicitly depends on the unknown θ. The details depend on the objectives of the analysis: if (X, Y) is observable and the goal is inference on θ, then we proceed as described in Martin and Liu (2013, 2015a); if only X is observable and the goal is prediction of Y, then we proceed as in Martin and Lingham (2016) or Cella and Martin (2020).

For the case of prediction, the idea is as follows. Given $X = x$, define a set-valued mapping $u \mapsto Q_x(u)$, into the space $\mathbb{Y} \times \Theta$ of unknown quantities, as

$$Q_x(u) = \{(y, \vartheta) \in \mathbb{Y} \times \Theta : (x, y) = a(\vartheta, u)\}.$$

If u satisfies the equation (5) with $X = x$, then $Q_x(u)$ contains the correct prediction. It is impossible to know for sure which u values satisfy the equation, but it is possible—since the distribution P_U is known—to construct a random set $\mathcal{U}$ of u values that we believe is likely to contain a solution. For such a $\mathcal{U}$, the new random set

$$Q_x(\mathcal{U}) = \bigcup_{u \in \mathcal{U}} Q_x(u),$$

obtained by mapping through the association to the space of unknowns, is equally likely to contain the correct prediction. Then we can define the lower and upper probabilistic predictor for Y, given $X = x$,

$$\begin{aligned} \underline{\pi}_x(A) &= \mathsf{P}_{\mathcal{U}}\{Q_x(\mathcal{U}) \subseteq A \times \Theta\} \\ \overline{\pi}_x(A) &= \mathsf{P}_{\mathcal{U}}\{Q_x(\mathcal{U}) \cap (A \times \Theta) \neq \varnothing\}, \end{aligned}$$

where $\mathsf{P}_{\mathcal{U}}$ is the distribution of the random set $\mathcal{U}$ and A is an arbitrary subset of $\mathbb{Y}$. The appropriate choice of random set $\mathcal{U}$ is beyond the scope of this short note, but suffice it to say that choosing $\mathcal{U} \sim \mathsf{P}_{\mathcal{U}}$ to achieve the validity condition is relatively straightforward; see Martin and Liu (2013, 2015a).

The above lower and upper prediction probabilities are belief and plausibility functions, respectively, defined on the power set of $\mathbb{Y}$, determined by the association, data, and user-defined random set. Our focus is on validity and efficiency, so we don't obligate ourselves to manipulating these functions using the Dempster–Shafer calculus of belief functions (Shafer, 1976; Dempster, 2008). Instead, the focus is on expressing the association between data and unknowns in terms of an auxiliary variable whose dimension is as small as possible. When the dimension is lower, the size of the random set needed to achieve validity is smaller, hence greater efficiency. General strategies for reducing the dimension were presented in Martin and Liu (2015b,c). The marginalization techniques in particular will be used below.

4.2 Three prisoners

For an IM solution, start with an association

$$Y = U_1$$
$$X = f(\theta, U_1, U_2),$$

where $U_1 \sim \text{Unif}(\{\text{A}, \text{B}, \text{C}\})$ and $U_2 \sim \text{Unif}(0,1)$ are independent, and

$$f(\theta, u_1, u_2) = \begin{cases} \text{B} & \text{if } u_2 \leq 1_{u_1=\text{C}} + \theta \, 1_{u_1=\text{A}} \\ \text{C} & \text{otherwise.} \end{cases}$$

A unique feature of this problem is that the quantity of interest, Y, the identity of the pardoned prisoner, has a known marginal distribution.

Since θ is not of primary interest, there is an opportunity to potentially reduce the auxiliary variable dimension before carrying out the IM construction (Martin and Liu, 2015c). Indeed, it is easy to check that, for every (x, y, u_2), there exists a θ such that $x = f(\theta, y, u_2)$. By the general IM marginalization theory, this implies the second equation in the association can be *effectively* ignored. This means valid (and efficient) prediction of Y should proceed based on its known marginal distribution. We say the second equation can be "effectively" ignored because it wouldn't make sense to predict that, say, $Y = \text{B}$ if we observe $X = \text{B}$. So we should account for this information in some way.

Based on the argument above, the A-step concludes by writing $Y = U$, where $U \sim \text{Unif}(\{\text{A}, \text{B}, \text{C}\})$. For the P-step, we introduce a suitable random set $\mathcal{U} \sim \text{P}_{\mathcal{U}}$ targeting the unobserved value of U. There are many options, but here we recommend to take $\mathcal{U}$ with support $\big\{\{\text{B}, \text{C}\}, \{\text{A}, \text{B}, \text{C}\}\big\}$ and masses assigned as

$$\text{P}_{\mathcal{U}}(\mathcal{U} = \{\text{B}, \text{C}\}) = \tfrac{2}{3} \quad \text{and} \quad \text{P}_{\mathcal{U}}(\mathcal{U} = \{\text{A}, \text{B}, \text{C}\}) = \tfrac{1}{3}.$$

With this choice, the probabilistic predictor returned by the IM's C-step is precisely the one described at the end of Section 3.1, the one that is valid and most efficient, superior to all the solutions presented in Gong and Meng (2020) based on updating the imprecise model according to formal rules.

4.3 Boxer, wrestler, and coin

For an IM solution, define an association as

$$Y_1 = 1_{U_1 \leq 0.5} \quad \text{and} \quad Y_2 = 1_{U_2 \leq \theta_{1|1}, U_1 \leq 0.5} + 1_{U_2 \leq \theta_{1|0}, U_1 > 0.5},$$

with $X = |Y_1 - Y_2|$ and (U_1, U_2) a pair of independent $\text{Unif}(0,1)$ random variables. Suppose, for example, that $X = 0$ is observed, i.e., that the outcomes of the fight and coin flip are the same; the case with $X = 1$ is analogous. When X is observed, the outcome of the fight determines the coin flip, and vice versa, so there's no need to consider both Y_1 and Y_2 after X is observed. We start with the case of Y_2, the fight's outcome. A generic (u_1, u_2) is pushed through the assertion, with $X = 0$, to a set in the (y_2, θ)-space:

$$Q_0(u_1, u_2) = \begin{cases} \{(1, \theta) : u_2 \leq \theta_{1|1}\} & \text{if } u_1 \leq 0.5 \\ \{(0, \theta) : u_2 > \theta_{1|0}\} & \text{if } u_1 > 0.5. \end{cases}$$

Since we're only interested in Y_2, our assertions about (Y_2, θ) take the form $\{y_2\} \times [0,1]^2$, for $y_2 \in \{0,1\}$. We'll leave out the details here, but it can be shown that, for any suitable random set $\mathcal{U} \subseteq [0,1]^2$, the probabilistic predictor for Y_2 returned by the IM is vacuous, i.e., its lower and upper probabilities are 0 and 1, respectively. As we showed above, this is the only valid solution.

Finally, if interest was in predicting Y_1, the outcome of the coin flip, then we could proceed very much like in the three prisoners example. That is, the general theory of marginal inference in Martin and Liu (2015c) allows us to ignore everything except Y_1, hence valid and efficient inference is achieved by using the marginal distribution of Y_1 to construct a valid and efficient probabilistic predictor. This agrees with the solution based on Dempster's rule and is more efficient than that based on the generalized Bayes rule.

5. CONCLUSION

The examples in Gong and Meng's paper are simultaneously both simple and challenging, making them ideal cases to test our understanding and to highlight the benefits of our perspective that focuses specifically on the construction of data-dependent beliefs that are both valid and efficient. This note is already too long, so we'll present our IM analysis of Simpson's paradox elsewhere.

It's interesting to see that, at least in cases where the imprecise model is known to be correctly specified, generalized Bayes is valid. But even in these relatively simple examples, we find that the IM solution can lead to more efficient prediction. In more complex settings, there the generalized Bayes solution faces certain challenges, in particular, specifying an imprecise model that is both sufficiently flexible and simple enough to compute the lower/upper envelopes. So there are ample reasons to consider alternative solutions. For example, Cella and Martin (2020) established a connection between valid IMs and the powerful conformal prediction machinery (Vovk, Gammerman and Shafer, 2005).

Finally, as we were preparing this discussion piece, it occurred to us that the failure of Fisher's fiducial argument and Dempster's extension thereof to achieve valid inference and prediction in general could possibly be understood in terms of the contraction, dilation, and/or sure loss examined by Gong and Meng. This claim, too, will be investigated further and our results will be presented elsewhere.

REFERENCES

CELLA, L. and MARTIN, R. (2020). Strong validity, consonance, and conformal prediction. arXiv:2001.09225.

DEMPSTER, A. P. (2008). The Dempster–Shafer calculus for statisticians. *Internat. J. Approx. Reason.* **48** 365–377. MR2419025

GONG, R. and MENG, X.-L. (2020). Judicious judgment meets unsettling updating: Dilation, sure loss, and Simpson's paradox. *Statist. Sci.*, to appear, arXiv:1712.08946.

MARTIN, R. and LINGHAM, R. T. (2016). Prior-free probabilistic prediction of future observations. *Technometrics* **58** 225–235. MR3488301

MARTIN, R. and LIU, C. (2013). Inferential models: a framework for prior-free posterior probabilistic inference. *J. Amer. Statist. Assoc.* **108** 301–313. MR3174621

MARTIN, R. and LIU, C. (2015a). *Inferential Models: Reasoning with Uncertainty*. *Monographs on Statistics and Applied Probability* **147**. CRC Press, Boca Raton, FL. MR3618727

MARTIN, R. and LIU, C. (2015b). Conditional inferential models: combining information for prior-free probabilistic inference. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 195–217. MR3299405

MARTIN, R. and LIU, C. (2015c). Marginal inferential models: prior-free probabilistic inference on interest parameters. *J. Amer. Statist. Assoc.* **110** 1621–1631. MR3449059

SHAFER, G. (1976). *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, N.J. MR0464340

VOVK, V., GAMMERMAN, A. and SHAFER, G. (2005). *Algorithmic Learning in a Random World*. Springer, New York. MR2161220

WALLEY, P. (1991). *Statistical Reasoning with Imprecise Probabilities*. *Monographs on Statistics and Applied Probability* **42**. Chapman & Hall Ltd., London. MR1145491

COMMENT ON GONG AND MENG

GREGORY WHEELER
FRANKFURT SCHOOL OF FINANCE & MANAGEMENT
g.wheeler@fs.de

forthcoming in STATISTICAL SCIENCE

Professors Gong and Meng's lucid and thought-provoking article views imprecise probability through the lens of three updating rules, highlighting discrepancies in inference between the *generalized Bayes rule*, on the one hand, and *Dempter's rule* and its dual, the *geometric rule*, on the other. In doing so, Gong and Meng vividly illustrate two important points, namely (*i*) inferential anomalies involving imprecise probabilities ought to be viewed as a helpful warning sign that some structural uncertainty looms in one's model, and (*ii*) such uncertainty is different in kind to sampling variability and therefore not resolved by updating with additional data.

Even so, the route Gong and Meng take to arrive at these two conclusions risks leaving the impression that the theory of imprecise probability is wobblier than it is. Specifically, in writing that,

> "in the world of imprecise probabilities, not only must we live with imperfections, but also accept intrinsic contradictions",

Gong and Meng suggest little has changed from the days of C.A.B. Smith's outline for inference with lower and upper personal "pignic odds" (Smith 1961), a proposal that Savage and de Finetti deemed "not fit for characterizing a new, weaker kind of coherent behaviour" (de Finetti and Savage 1962).

In my remarks, I would like to offer a corrective to the notion that inference with imprecise probabilities is plagued by inherent contradictions. On the contrary, for the contemporary theory of *lower previsions* (Walley 1991; Troffaes and de Cooman 2014), which includes lower probabilities as a special case, coherence preservation under inference is inviolable. Yet, once sure-loss avoidance is promoted to a fundamental principle, both Dempster's rule and the geometric rule fall by the wayside—except in specific, benign circumstances where their application is guaranteed to avoid sure loss.

1 SURE LOSS AVOIDANCE & COHERENCE

Whether to accept sure-loss avoidance as fundamental will depend on what you get from the theory of lower previsions in return. Gong and Meng rightly observe that if lower and upper previsions are interpreted as acceptable one-sided betting odds, with lower previsions denoting the maximum buying price you would pay for a gamble and upper previsions denoting the minimum selling price you would accept for that gamble, then it is natural to accept sure-loss avoidance as a principle of rationality. They nevertheless contrast this *direct* interpretation of a lower prevision, as a representation of your disposition to bet on a collection of gambles, with an *indirect* interpretation that regards a lower prevision as a summary of the set of probabilities that are compatible with an incompletely specified model. This indirect interpretation is central to Bayesian sensitivity analysis, but it has also played an important role in the historical development of imprecise probabilities more generally.

For instance, Smith showed that every coherent lower prevision may be understood as the lower envelope of some set of linear previsions, a result that Walley later strengthened to a characterization (1991, §3.3): specifically, a lower prevision $\underline{P}$ *avoids sure loss* if and only if there is a linear prevision P such that $P(X) \geq \underline{P}(X)$, for all gambles X on a fixed domain, and $\underline{P}$ is a *coherent lower prevision* if and only if there is a set of linear previsions $\mathbb{P}$ such that $\underline{P}$ is the lower envelope of $\mathbb{P}$, that is $\underline{P}(X) = \inf\{P(X) : P \in \mathbb{P}\}$, for all X on a similarly shared domain. When the range of X is restricted to $\{0,1\}$, X works as an indicator function and $\underline{P}(X)$ as a lower probability. Such sure-loss avoidance and coherence conditions extend to conditional lower previsions, too.

The question then is whether the inferential capabilities that one would need when approximating a true but unknown probability distribution can be subsumed under the machinery developed for lower previsions based on a direct, behavioral interpreta-

tion. Walley argued that it does (1991, §2.10) and I agree, with one qualification.

That qualification, a benefit of hindsight, is to concede that managing coherence conditions for conditional lower previsions is complicated when those conditions are tied to a set of linear previsions in the (customary) manner sketched above. One reason why is that the familiar equivalence between additive probability and linear previsions does not carry over to lower probability and lower previsions. A linear prevision is simply the expectation calculated by taking the integral with respect to a given probability, and this equivalence licenses Bayesians to treat "degrees of belief" expressed over a language of events as fundamental. However, an analogous one-to-one correspondence between lower probability and lower previsions does not hold. Specifically, unlike linear previsions, two lower previsions can agree in values for all *events*, and therefore express the same lower probabilities, but still express different values over *gambles*. This one-to-many relationship means that commonplace probabilistic intuitions can go haywire in the context of imprecise probabilities, resulting in some forms of reasoning that are valid for precise probabilities being invalid for imprecise probabilities.

Since Walley's chef-d'œuvre, simpler and more unified inference methods for conditional lower previsions have been developed (Troffaes and de Cooman 2014), but they have come about by abandoning the notion that sets of probabilities are elemental. Whereas the Old Testament approach to imprecise probabilities closely links lower previsions to sets of probabilities, thereby setting a difficult path for coherent inference to follow, the New Testament puts coherence and inference first but demotes (closed convex) sets of probabilities to derivable or dispensable representations, as need be. Instead, *desirable* or *acceptable gambles* are treated as fundamental, where a gamble X on a set of possibilities is a real-valued map from those possibilities, interpreted as the gain or loss that you associate with each possible state. Then, $\underline{P}(X)$ represents the supremum price you are willing to pay in exchange for the gamble X, and a conditional lower prevision of the gamble X given the $\{0,1\}$-gamble G, $\underline{P}(X|G)$ is your lower prevision for X contingent on the event G occurring ($G=1$), which is "called-off" otherwise ($G=0$).

Briefly, and to just give a flavor, there are four simple yet constructive axioms for a coherent set $\mathbb{D}$ of desirable gambles. The first two, which are rationality axioms, mandate that you ought to (*i*) never accept a gamble you cannot win (i.e., do not include in $\mathbb{D}$ an X whose vector of values is everywhere negative), and (*ii*) always accept a gamble you cannot lose. The 0-gamble denotes *status quo ante*, and there are variants of these axioms which include, rather than exclude, 0-gambles among a coherent set of gambles—a difference reflected, even if only loosely observed, in the terminology used to refer to the strict desirability of gambles or merely to their acceptability. The second pair of axioms are closure conditions, encoding the properties of a linear scale for evaluating gambles, namely (*iii*) positive scale invariance and (*iv*) a combination rule whereby if X and Y are each acceptable gambles, then $X+Y$ ought to be acceptable to you, too.

The generalized Bayes rule in this scheme is simply

$$\underline{P}(G[X-\underline{P}(X|G)])=0 \qquad (1)$$

where it is assumed that both $\underline{P}(G)>0$ and the contingent gamble $G[X-\underline{P}(X|G)]$ are in $\mathbb{D}$. Methods for conditioning and updating on zero-probabilities have been simplified, too (De Bock and de Cooman 2015).

2 What price for generality?

The New Testament's full embrace of modeling uncertainty in terms of the rationality of beliefs and behavioral dispositions might appear to go too far, even among those who otherwise favor the Bayesian approach. Yet, the contemporary theory of lower previsions is a general framework attuned to foundational issues of the kind that Gong and Meng raise, and as such includes traditional linear previsions as a special case, much like first-order logic includes propositional logic as a special case. Lower previsions offer an alternative way of conceiving and working with probability models, not an alternative to probability altogether. Viewed in this light, it is perhaps less surprising to find that sets of probabilities are derivable from, rather than foundational to, lower previsions.

The analogy to logic goes a bit further. Consider some differences between propositional logic, which dates back two millennia, and first-order logic, which is just over a century old. Both the syntax and semantics of first-order logic work very differently than the syntax and semantics of propositional logic. First-order logic admits syntactically well-formed "open" sentences which are nevertheless uninterpretable, semantically, until "closed"

under quantification. There is no such thing as a syntactically well-formed formula of propositional logic that is semantically uninterpretable, however. Every formula of propositional logic is interpreted by evaluating all logically exhaustive combinations of its interpretations, such as may be displayed in a truth table, which is impossible to do for first-order logic. As for inference, propositional logic is decidable whereas first-order logic is not. Yet, if one were to maintain that semantic interpretability and syntactic well-formedness were inseparable properties of logical formulas, truth tables fundamental to model theory, or decidability essential to logic itself, the world of first-order logic would be regarded as imperfect and contradictory, too. We generally don't take that view, however, and similar slack should be afforded to lower previsions—or so I would argue. Space prohibits more than a gesture here, but a paper-sized treatment appears elsewhere (Wheeler 2021).

The main point is this. Trouble for imprecise probabilities rarely comes in the form of inherent contradictions, but instead is more apt to arise from seeking to preserve consistency at all costs. Disjunction, for instance, is missing from the vocabulary of desirable gambles, and is tricky to deal with. Recent work using desirable gamble *sets* to construct choice functions (De Bock and de Cooman 2019) offers a promising avenue to address this deficiency, however. This extension offers the capability to say of a set of gambles that at least one is desirable without necessarily identifying which it is. Accommodating set-based choice also suggests a means, in a coherence preserving setting, to address problems of the kind that motivate the use of belief functions.

3 Dilation and Association

Which brings us to dilation. Dilation occurs when the interval estimate of an event E is properly included in the interval estimate of E conditional on every element of some measurable partition $\mathcal{B}$. As Gong and Meng point out, in such cases, updating by the generalized Bayes rule on *any* value of $\mathcal{B}$ would render your initial estimate of E less precise. Should you update or instead refuse information that would resolve your uncertainty about $\mathcal{B}$? Would you be willing to pay some amount, however small, to remain ignorant? With dilation, one could be forgiven for thinking, *so much for consistency*.

Yet, the notion that you can be better off with less information is not unheard of in the theory of games. Akerlof's study of market failures in the used car market, circa 1970, is a prime example. A customer will not know, but a used-car salesman will know, which cars on the lot are lemons. Wary of being fleeced, a customer will refuse to pay more than the going rate for a bad car, if not refuse to trade altogether. For the salesman then there is a disadvantage knowing more about the quality of the cars on the lot than his customers do, as no car, good or bad, can command a good-car price.

Akerlof's demonstration of *adverse selection* is an example of a strategic interaction in which information asymmetry backfires on the player with more information. Some textbook treatments of adverse selection maintain that negative-valued information cannot occur in single-person decision problems, however, as act-state independence would rule out the type of act-state dependence that the customer on the car lot fears will be used against him. But this is only true for single-person decision problems with additive probabilities. Dilation illustrates a form of state uncertainty, which lower previsions capture, that is sufficient to break the independence condition that ordinary decision problems take for granted. Put more carefully, dilation examples do not explicitly rule out that the pair of events in question are dependent. And a cleverly designed dilation example will prey on intuitions that are misleading in an imprecise probabilities setting, particularly those to do with structural properties of independence and association.

At root, dilation is not so much an updating paradox as a result of reasoning as if stochastic independence holds when it does not. Although Gong and Meng remark that "generalized notions of association and independence...are yet to be defined for sets of probabilities", there are several logically distinct notions of independence for imprecise probabilities (Couso, Moral, and Walley 1999). Here reference to an explicit sets of probabilities helps. For instance, for an ordinary additive probability $p \in \mathbb{P}$ and events A, B, you know that if B is irrelevant to A with respect to this p, that is, if $p(A|B) = p(A)$, then A is irrelevant to B, and the joint distribution of A and B is the product of the pair of marginal distributions. But each step in this sequence of valid inferences is invalid for imprecise probabilities. Irrelevance for lower previsions is not symmetric, and even when both A is irrelevant to B and B irrelevant to A, the set of joint distributions might not factorize. The converse of each is valid, however,

pointing to a range of strong to weak independence concepts.

Wily dilation examples are often constructed to satisfy weaker notions of irrelevance without satisfying full, factorized stochastic independence, and will in fact include a distribution in $\mathbb{P}$ for which the pair of events are positively associated and another distribution for which they are negatively associated (Pedersen and Wheeler 2014), an observation that is easily adapted to include asymmetric cases in which one event dilates another but not vice versa (Pedersen and Wheeler 2019).

But if this explains what dilation is, what should be done about it? I agree with Gong and Meng that dilation alone is not a problem, anymore than an open formula of first-order logic is itself a problem. But instead of opting for an alternative updating rule, and braving the hazards they bring, I prefer to stick to the generalized Bayes rule and simply select an appropriate decision rule. In fact, returning to the questions above that suggest you might be rationally compelled by dilating probabilities to either ignore information or even pay someone to avoid it, such injunctions depend crucially on your choice of decision rule. In fact, some decision rules for imprecise probabilities preserve the principle that no decision maker should be made worse off, in expectation, from receiving free information (Pedersen and Wheeler 2015).

To be fair, a remnant of the updating anomalies that Gong and Meng discuss carries over to decision making with imprecise probabilities. There is no single decision rule that is unequivocally best, and the current state of the art is far less tidy. A complaint might then be lodged that this only kicks the inference can down the pick-the-right-decision-rule road, and there is a kernel of truth to this. But, that is a discussion to save for another day.

In closing, I commend Gong and Meng for their valuable contribution and wish to stress once more how much I agree with them in the main. Lower previsions afford much greater expressive capacity and, as a consequence, pull apart some notions that are unitary concepts in standard, additive probability models. Thus, it is a natural response to view updating anomalies like dilation as a helpful pointer to some of the novel implications that follow from uncertainty in such settings.

References

Couso, I., S. Moral, and P. Walley (1999). Examples of independence for imprecise probabilities. In G. de Cooman (Ed.), *Proceedings of the First Symposium on Imprecise Probabilities and Their Applications (ISIPTA)*, Ghent, Belgium.

De Bock, J. and G. de Cooman (2015). Conditioning, updating and lower probability zero. *International Journal of Approximate Reasoning 67*, 1–36.

De Bock, J. and G. de Cooman (2019). Interpreting, axiomatising and representing coherent choice functions in terms of desirability. In J. De Bock, C. P. de Campos, G. de Cooman, E. Quaeghebeur, and G. Wheeler (Eds.), *Proceedings of the Eleventh International Symposium on Imprecise Probabilities: Theories and Applications*, Volume 103 of *Proceedings of Machine Learning Research*, Thagaste, Ghent, Belgium, pp. 125–134.

de Finetti, B. and L. J. Savage (1962). Sul modo di scegliere le probabilità iniziali. *Biblioteca del Metron, Serie C 1*, 81–154.

Pedersen, A. P. and G. Wheeler (2014). Demystifying dilation. *Erkenntnis 79*(6), 1305–1342.

Pedersen, A. P. and G. Wheeler (2015). Dilation, disintegrations, and delayed decisions. In *Proceedings of the 9th Symposium on Imprecise Probabilities and Their Applications (ISIPTA)*, Pescara, Italy, pp. 227–236.

Pedersen, A. P. and G. Wheeler (2019). Dilation and asymmetric relevance. In J. De Bock, C. P. Campos, G. de Cooman, E. Quaeghebeur, and G. Wheeler (Eds.), *Proceedings of Machine Learning Research*, Volume 103 of *Proceedings of the 11th Symposium on Imprecise Probabilities and Their Applications (ISIPTA)*, pp. 324–326.

Smith, C. A. B. (1961). Consistency in statistical inference (with discussion). *Journal of the Royal Statistical Society 23*, 1–37.

Troffaes, M. C. M. and G. de Cooman (2014). *Lower Previsions*. Chichester, West Sussex: Wiley and Sons.

Walley, P. (1991). *Statistical Reasoning with Imprecise Probabilities*. London: Chapman and Hall.

Wheeler, G. (2021). A gentle approach to imprecise probability. In T. Augustin, F. Cozman, and G. Wheeler (Eds.), *Reflections on the Foundations of Statistics: Essays in Honor of Teddy Seidenfeld*, Theory and Decision Library A. Springer.

On Focusing, Soft and Strong Revision of Choquet Capacities and Their Role in Statistics: Comments on Gong and Meng's Paper

Thomas Augustin and Georg Schollmeyer

Abstract. We congratulate Ruobin Gong and Xiao-Li Meng on their thought-provoking paper demonstrating the power of imprecise probabilities in statistics. In particular, Gong and Meng clarify important statistical paradoxes by discussing them in the framework of generalized uncertainty quantification and different conditioning rules used for updating. In this note, we characterize all three conditioning rules as envelopes of certain sets of conditional probabilities. This view also suggests some generalizations that can be seen as compromise rules. Similar to Gong and Meng, our derivations mainly focus on Choquet capacities of order 2, and so we also briefly discuss in general their role as statistical models. We conclude with some general remarks on the potential of imprecise probabilities to cope with the multidimensional nature of uncertainty.

Key words and phrases: Imprecise probabilities, Choquet capacities, Updating, Neighborhood models, Generalized Bayes rule, Dempster's rule of conditioning.

1. INTRODUCTION

In their stimulating paper "Judicious Judgment Meets Unsettling Updating: Dilation, Sure Loss, and Simpson's Paradox", Ruobin Gong and Xiao-Li Meng (hereafter GM) offer a fresh perspective on famous problems that have long shaken the foundations of statistical analysis. GM manage to trace the paradoxes back to self-contradictory model assumptions about the marginals and the joint distribution and creatively relate them to phenomena occurring in updating imprecise probabilities. These insights are an excellent example of how the general framework of imprecise probabilities, through its expanded understanding of uncertainty, not only provides new opportunities for statistical modeling, but also helps to illuminate hidden implicit assumptions in classical modeling.

In this short note, we provide in Section 2 some variations of the central topic of conditioning under a generalized probabilistic setting. We will make explicit some mathematical properties of Choquet capacities of order 2 that are contained implicitly in GM's paper. In particular, these properties will allow us to characterize the three different ways of conditioning as envelopes of certain sets of conditional probabilities. In the light of this characterization, we will revisit the notions of "being cautious" and "overfitting", contrasting the generalized

Thomas Augustin is Professor of Statistics and Head of the Foundations of Statistics and their Applications Group at the Department of Statistics, Ludwigs-Maximilians Universität München (LMU Munich), Germany. (e-mail: thomas.augustin@stat.uni-muenchen.de). Georg Schollmeyer is a post-doctorial staff member there. (e-mail: georg.schollmeyer@stat.uni-muenchen.de).

Bayes rule (GBR) and Dempster's rule as extreme positions that also allow generalizations by taking an intermediate position. In Section 3 we will address the question of how general the assumed model class of Choquet capacities of order 2 actually is, and thus which practically relevant models are covered by it. Section 4 is reserved for some general concluding remarks on the potential of imprecise probabilities in the context of complex uncertainty.

2. ENVELOPE REPRESENTATIONS OF THE DIFFERENT CONCEPTS OF CONDITIONAL PROBABILITIES

2.1 A Common Representation

Our argumentation below strongly relies on the following lemma, guaranteeing that for Choquet capacities of order 2 the lower respectively upper probabilities $\underline{P}$ and $\overline{P}$ of chains of events are *simultaneously* attained by a classical probability in the induced set of compatible distributions. (For further reference and in accordance with the literature, we use the term *credal set* (induced by $\underline{P}$ and $\overline{P}$) for this set of compatible distributions in the set $\mathcal{M}$ of all distributions on the considered measurable space.)

LEMMA 1. [1] *Let $\underline{P}$ be a lower probability such that its credal set $\mathcal{P} = \{P \in \mathcal{M}, \underline{P} \geq P\}$ is relatively compact. Then $\underline{P}$ is two-monotone if and only if for every chain of events $E_1 \subseteq E_2 \subseteq E_3 \ldots \subseteq E_n$ there exists a probability $P \in \mathcal{P}$ such that $P(E_i) = \underline{P}(E_i)$ for all $i \in \{1, \ldots, n\}$.*

This lemma is used in GM's paper implicitly, for instance, in the closed-form reformulation of the generalized Bayes rule in (GM, 2.11f) valid for Choquet capacities of order 2. Using it explicitly, and applying it to the events $E_1 = A \cap B$ and $E_2 = B$, shows that the ratios in (GM, 2.8), and in (GM, 2.9) respectively, are simultaneously optimized. Assuming $\underline{P}(B) > 0$ to make all expressions well-defined, this allows to rewrite the considered types of conditional probabilities in a unified way (cf., e.g., Gilboa and Schmeidler (1993)).

$$(1) \qquad \underline{P}_{\mathfrak{Z}}(A|B) = \inf_{P \in \mathcal{P}_{\mathfrak{Z}}} \frac{P(A \cap B)}{P(B)} \qquad \text{and} \qquad \overline{P}_{\mathfrak{Z}}(A|B) = \sup_{P \in \mathcal{P}_{\mathfrak{Z}}} \frac{P(A \cap B)}{P(B)},$$

where

$$(2) \quad \mathcal{P}_{\mathfrak{Z}} = \begin{cases} \{P \in \mathcal{M} | p \geq \underline{P}\} \stackrel{def}{=} \mathcal{P} & \text{Gen. Bayes Rule} \\ \{P \in \mathcal{M} | P \geq \underline{P} \wedge P(B) = \overline{P}(B)\} \stackrel{def}{=} \mathcal{P}_{\mathfrak{D}} \text{ in the case of} & \text{Dempster's Rule} \\ \{P \in \mathcal{M} | p \geq \underline{P} \wedge P(B) = \underline{P}(B)\} \stackrel{def}{=} \mathcal{P}_{\mathfrak{G}} & \text{Geometric Rule} \end{cases}$$

2.2 Focusing versus (Strong) Belief Revision

The envelope representation illustrates GM's important distinction between two different conceptualizations of updating, namely updating as belief revision versus updating as focusing (cf., Dubois and Prade (1997)). In focusing, generic knowledge is not changed, instead, it is only applied to the event that corresponds to the observed data. This leads to the generalized Bayes rule. In contrast, in belief revision one modifies generic knowledge or factual evidence about a problem in the light of new knowledge or evidence. Equation (2) underlines that both the geometric rule as well as Dempster's rule perform a rather strong revision, which may also be interpreted as a strong "overfitting". Constructing $\mathcal{P}_{\mathfrak{D}}$ and $\mathcal{P}_{\mathfrak{G}}$, they both rely exclusively on a single value taken from the interval $[\underline{P}(B), \overline{P}(B)]$. While the geometric rule confines itself on the lowest value, Dempster's rule concentrates on the highest one.[2] In a classical

[1] For a proof, see, e.g., Chateauneuf and Jaffray (1989, Proposition 12, p. 277).

[2] This argumentation understands, in accordance with GM's paper, Dempster's rules of conditioning and combination as producing a non-additive set-function enveloping a set of probabilities. To avoid misunderstandings, it may be noted explicitly that in the so-called *Dempster-Shafer Theory of Belief Functions* popular in artificial intelligence this interpretation is strongly rejected by many authors: "Most important, a probability-bound interpre-

parametric Bayesian setting, where the prior distribution of a parameter ϑ is updated, based on sample B, to the corresponding posterior distribution, $P(B)$ is the predictive distribution of the sample. Then, Dempster's rule refines the underlying credal set $\mathcal{P}$ to contain only those probabilities giving the sample the highest likelihood. Indeed, Gilboa and Schmeidler (Gilboa and Schmeidler (1994, 1993), see also, Dubois and Prade (1997)) denote Dempster's rule as "maximum likelihood update". Moreover, in particular if we understand $\mathcal{P}$ as parameterized by a nuisance parameter, Dempster's rule can be interpreted as an empirical Bayes approach. It corresponds to the so-called *ML-II approach* (e.g., Berger, 1985, Section 3.5.4), originally suggested by Good (see Good (1983, e.g., p. 46f)).

In this sense, one can also conceptually differentiate the generalized Bayes rule and Dempster's rule as an ideal type dichotomy between an optimistic view and a pessimistic/conservative view. While according to the generalized Bayes rule the conditional lower probability is obtained as the worst conditional classical probability that is consistent with the given lower and upper probabilities, Dempster's rule can be viewed as a very optimistic approach, radically excluding all probability functions that are not maximally plausible in the light of the observed event B. Somewhere within (and to some extent also somewhere beside?) this ideal type dichotomy, the geometric rule can be located as a rule, which, in contrast to the GBR, restricts $\mathcal{P}$ for updating, however, in contrast to Dempster's rule, in a pessimistic way. It restricts $\mathcal{P}$ to all compatible probabilities that assign the lowest possible probability to the observed event B. Although somehow parallel in construction to Dempster's rule, this way of restricting $\mathcal{P}$ by relying on the lowest possible likelihood is a minimax perspective, that is very cautious from the learning point of view. Indeed, quite naturally, this rule, can not sharpen vacuous prior information (compare Section 4.3 in GM's paper).

2.3 Soft Revision and Likelihood Cuts

These deliberations suggest a quite natural compromise between optimism and pessimism, between the conservative focusing on one hand and the strong revision of Dempster's rule (and the geometric rule) on the other hand, which can be suspected to posses a strong tendency towards overfitting. Instead of basing the revision on one of the interval limits of $[\underline{P}(B), \overline{P}(B)]$, one relies on a subinterval of high or small values. More concretely, for a fixed real value $\alpha \in [0,1]$, one replaces in (2) the condition $P(B) = \overline{P}(B)$ by the condition[3]

$$P(B) \geq \alpha \cdot \overline{P}(B)\,, \tag{3}$$

or dually, the condition $P(B) = \underline{P}(B)$, which is equivalent to $P(B^c) = \overline{P}(B^c)$), by

$$P(B^c) \geq \alpha \cdot \overline{P}(B^c) \tag{4}$$

to obtain suitable generalizations of Dempster's rule and the geometric rule, respectively.[4] For $\alpha = 0$ we obtain GBR, and for $\alpha = 1$ we reproduce Dempster's rule or the geometric rule, respectively. In this sense, α can be seen here as a 'parameter of revision'. For a small, but positive value α, these revision rules do not rigidly revise the model to only the compatible probabilities that give the observed event B the most/least probability. Such soft revisioning rules may be quite attractive when one feels uncomfortable with the overfitting character of strong revision rules.

Soft revision rules are not coherent in the sense of Walley (1991)'s general coherence theory justifying the GBR. In fact, the GBR does not perform any revisioning at all; it never

tation is incompatible with Dempster's rule for combining belief functions. If we make up numbers by thinking of them as lower bounds on true probabilities, and we then combine these numbers by Dempster's rule, we are likely to obtain erroneous and misleading results." Shafer (1990, p. 335) Then, belief functions derived from Dempster's rule of conditioning, and more generally from Dempster's rule of combination, are understood as providing an uncertainty calculus of its own. (For a recent review see Denoeux (2016).)

[3] This approach has already been introduced by Cattaneo (2014).

[4] Another variant of generalization would be to replace $P(B) = \overline{P}(B)$ by $P(B) \geq \underline{P}(B) + \alpha \cdot (\overline{P}(B) - \underline{P}(B))$ and to replace $P(B) = \underline{P}(B)$ analogously. Other generalizations are, of course, thinkable as well, for instance neighborhood-models around the maximizing/minimizing probabilities. As a further alternative, Held, Augustin and Kriegler (2008) consider a mixture of the layers produced by the different values of $[\underline{P}(A), \overline{P}(A)]$.

changes the priori assessments, but merely focuses on the implication for situations in which B is observed. As discussed in Section 4.3 of GM's paper, one can thus not learn with the GBR from vacuous prior knowledge.[5] In contrast, the α-cut rule with $\alpha > 0$ and congenial rules are able to learn from vacuous priors.

3. ON THE ROLE OF TWO-MONOTONE CHOQUET-CAPACITIES IN STATISTICS

Many of the results in GM's paper build on the condition that the lower and upper probabilities are Choquet capacities of order 2. In this section, we look at the natural question arising how restrictive this assumption is from a statistical modeling perspective.

From the principle point of view of the general theory of imprecise probabilities, the condition of two-monotonicity seems artificial. Neither in the behavioral approach to imprecise probabilities (see, in particular, Walley (1991)) nor within its frequentist counterpart (developed by Fine and students, e.g. Fierens, Rego and Fine (2009)), two-monotonicity has a contextual meaning or natural interpretation. In addition, two-monotonicity plays also no prominent role in the interpretation-independent branch of imprecise probabilities following Weichselberger (2001). Nevertheless, two-monotone lower probabilities are quite attractive for statistics. In particular, following the prominent Huber-Strassen Theorem (Huber and Strassen (1973), see also Augustin, Walter and Coolen (2014, Section 7.5.2) for a review of work building on it), two-monotone lower probabilities allow for a rigorous generalization of Neyman-Pearson tests to imprecise probabilities.

A very rich class of two-monotone lower probabilities, which historically also motivated the development of the Huber-Strassen theorem, is provided by certain neighborhood models (see, e.g., Augustin and Hable (2010); Montes, Miranda and Destercke (2020a,b)). They allow, quite attractively, to formalize the notion of "approximately true distributions", for instance, by considering all distributions close to a certain *central distribution* p^*. Therefore, neighborhood models have been used in particular in robust statistics as an imprecise sampling distribution or in robust Bayesianism as generalized prior distributions. Typical examples include the δ-total variation model, comprising all distributions where the total variation distance to p^* is smaller than δ, or the ϵ-contamination model formalizing the situation where at least $(1-\epsilon)\cdot 100\%$ of the observations follow the central distribution p^*, but the remaining $\epsilon \cdot 100\%$ may just follow any arbitrary distribution. Generally, many neighborhood models can be written in the form $f \circ p^*$, where convexity of f guarantees two-monotonicity.[6]

Other natural ways of constructing two-monotone models are discrete models with bounds on the probabilities of singletons only (*probability intervals*, Weichselberger and Pöhlmann (1990)) or bounds on distribution functions. The latter, often called *p-boxes*, play a prominent role in generalized uncertainty quantification in reliability analysis (see, e.g., Destercke, Dubois and Chojnacki (2008)).

Finally, also a natural connection between the granularity of observation and two-monotonicity shall be mentioned. Given two measurable spaces $(\Omega, \mathcal{A})$, and $(\Omega, \mathcal{F})$ with $\mathcal{F} \supseteq \mathcal{A}$, a two-monotone lower probability $\underline{P}^*$ can be constructed by extending a two-monotone lower probability $\underline{P}$ on $\mathcal{A}$ to events in $\mathcal{F}$ by natural extension (cf. Walley (1981, p. 52)), and the different concepts of conditioning can be applied. Naturally, if $\underline{P}$ is a classical probability and conditioning is performed by considering only partitions in $\mathcal{A}$, all considered concepts of conditioning coincide in this case.

[5] To guarantee that GBR-like inferences with vacuous priors lead to non-vacuous posteriors, extreme prior probabilities have to be excluded. This is achieved by the rather prominent *Imprecise Dirichlet Model* (Walley (1996)) for inference from categorical data. Generally, so-called *near-ignorance prior models* can be considered (see, in particular, Benavoli and Zaffalon (2014)'s approach for multivariate exponential families).

[6] Such models are also used in insurance mathematics as *distorted probabilities*, see, for instance, Wang and Young (1998) for premium calculation in this context, where also the GBR and Dempster's rule are discussed.

4. SOME GENERAL CONCLUDING REMARKS

From a principled and general perspective, we unanimously share GM's enthusiasm for a generalized understanding and modeling of uncertainty. What had become obvious in the first AI summer in the context of expert systems and general systems theory, is currently even more important in the environment of ubiquitous and widely available data."Uncertainty is a multidimensional concept. [However, its ...] multidimensional nature was obscured when uncertainty was conceived solely in terms of [classical] probability theory, in which it is manifested by only one of its dimensions." (Klir and Wierman, 1999, p. 1)

Indeed, as statisticians and data scientists, we have to pay more attention to the so-to-say "big data uncertainty", i.e. those dimensions of uncertainty that go beyond sampling uncertainty and thus do not vanish with increasing sample size. Only generalized probabilistic approaches used in a sophisticated way as in GM's paper allow to distinguish between variability and indeterminacy, which is crucial for an appropriate modeling of the quality of probabilistic information. These models are naturally imprecise, or – to avoid the unfortunate misnomer 'imprecise' for actually better and more accurate models – rather, set-valued. This set-valued character promises to express scarce, conflicting or simply incomplete information without having to rely on unwarranted assumptions. We agree with GM that making strong but untestable assumptions about unobservable structures just for the sake of a seemingly precise result undermines practical relevance of the statistical analysis, well aware of the "Law of Decreasing Credibility",

> "The credibility of inferences decreases with the strength of the assumptions maintained" (Manski, 2003, p. 1),

as Manski and his followers put it in the area of partial identification, a rather parallel running development of powerful set-valued analysis in econometrics.[7]

REFERENCES

AUGUSTIN, T. and HABLE, R. (2010). On the impact of robust statistics on imprecise probability models: A review. *Structural Safety* **32** 358–365.

AUGUSTIN, T., WALTER, G. and COOLEN, F. (2014). Statistical Inference. In *Introduction to Imprecise Probabilities* (T. Augustin, F. Coolen, G. de Cooman and M. Troffaes, eds.) 135–189. Wiley.

BENAVOLI, A. and ZAFFALON, M. (2014). Prior near ignorance for inferences in the k-parameter exponential family. *Statistics* **49** 1104–1140.

BERGER, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. Springer.

CATTANEO, M. E. G. V. (2014). A continuous updating rule for imprecise probabilities. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems* (A. LAURENT, O. STRAUSS, B. BOUCHON-MEUNIER and R. R. YAGER, eds.) 426–435. Springer.

CHATEAUNEUF, A. and JAFFRAY, J.-Y. (1989). Some characterizations of lower probabilities and other monotone capacities through the use of Möbius inversion. *Mathematical Social Sciences* **17** 263 - 283.

DENOEUX, T. (2016). 40 years of Dempster–Shafer theory. *International Journal of Approximate Reasoning* **79** 1–6.

DESTERCKE, S., DUBOIS, D. and CHOJNACKI, E. (2008). Unifying practical uncertainty representations – I: Generalized p-boxes. *International Journal of Approximate Reasoning* **49** 649 - 663.

DUBOIS, D. and PRADE, H. (1997). Focusing vs. belief revision: A fundamental distinction when dealing with generic knowledge. In *Qualitative and Quantitative Practical Reasoning* (D. M. GABBAY, R. KRUSE, A. NONNENGART and H. J. OHLBACH, eds.) 96–107. Springer, Berlin, Heidelberg.

FIERENS, P. I., REGO, L. C. and FINE, T. (2009). A frequentist understanding of sets of measures. *Journal of Statistical Planning and Inference* **139** 1879–1892.

GILBOA, I. and SCHMEIDLER, D. (1993). Updating ambiguous beliefs. *Journal of Economic Theory* **59** 33 - 49.

GILBOA, I. and SCHMEIDLER, D. (1994). Additive representations of non-additive measures and the Choquet integral. *Annals of Operations Research* **52** 43–65.

GOOD, I. (1983). *Good Thinking: The Foundations of Probability and Its Applications*. University of Minnesota Press, Minneapolis, MN (reprint 2009: Dover, Mineola, NY).

HELD, H., AUGUSTIN, T. and KRIEGLER, E. (2008). Bayesian learning for a class of priors with prescribed marginals. *International Journal of Approximate Reasoning* **49** 212–233.

[7] See also the surveys Molinari (2020) on the development of partial identification in micoreconometrics and Molchanov and Molinari (2018) on the use of random sets in the context of partial identification.

HUBER, P. and STRASSEN, V. (1973). Minimax tests and the Neyman-Pearson lemma for capacities. *The Annals of Statistics* **1** 251–263.

KLIR, G. J. and WIERMAN, M. J. (1999). *Uncertainty-Based Information. Elements of Generalized Information Theory*. Physica.

MANSKI, C. (2003). *Partial Identification of Probability Distributions*. Springer.

MOLCHANOV, I. and MOLINARI, F. (2018). *Random Sets in Econometrics*. Cambridge University Press.

MOLINARI, F. (2020). Microeconometrics with partial identification. In *Handbook of Econometrics. Volume 7, Part A* (S. N. Durlauf, L. P. Hansen, J. J. Heckman and R. L. Matzkin, eds.) 355–486.

MONTES, I., MIRANDA, E. and DESTERCKE, S. (2020a). Unifying neighbourhood and distortion models: Part I – new results on old models. *International Journal of General Systems* **49** 602–635.

MONTES, I., MIRANDA, E. and DESTERCKE, S. (2020b). Unifying neighbourhood and distortion models: Part II – new models and synthesis. *International Journal of General Systems* **49** 636–674.

SHAFER, G. (1990). Perspectives on the theory and practice of belief functions. *International Journal of Approximate Reasoning* **4** 323 - 362.

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

WALLEY, P. (1991). *Statistical Reasoning with Imprecise Probabilities*. Chapman & Hall, London.

WALLEY, P. (1996). Inferences from multinomial data: Learning about a bag of marbles. *Journal of the Royal Statistical Society. Series B* **58** 3-57.

WANG, S. S. and YOUNG, V. R. (1998). Risk-adjusted credibility premiums using distorted probabilities. *Scandinavian Actuarial Journal* **1998** 143-165.

WEICHSELBERGER, K. (2001). *Elementare Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung I: Intervallwahrscheinlichkeit als umfassendes Konzept [Elementary Foundations of a More General Calculus of Probability I: Interval Probability as a Comprehensive Concept]*. Physica, Heidelberg.

WEICHSELBERGER, K. and PÖHLMANN, S. (1990). *A Methodology for Uncertainty in Knowledge-Based Systems*. Springer, Heidelberg.

Submitted to Statistical Science

Rejoinder: Let's be imprecise in order to be precise (about what we don't know)

Ruobin Gong and Xiao-Li Meng

Preparing a rejoinder is a typically rewarding, sometimes depressing, and occasionally frustrating experience. The rewarding part is self-evident, and the depression sets in when a discussant has much deeper and crisper insights about the authors' thesis than authors themselves. Frustrations arise when the authors thought they have made some points crystal clear, but the reflections from the discussants show a very different picture. We are deeply grateful to the editors of *Statistical Science* and the discussants for providing us an opportunity to maximize the first, sample the second, and minimize the third.

1. LET'S ARGUMENT OUR SHOES TO FIT OUR GROWING FEET

The first discussion we received was from Professor Glenn Shafer, who kindly sent us a draft weeks before the submission deadline. We immediately knew that we would be in for a rare intellectual feast. His historically richly infused and theoretically deeply fermented insights provide us with an intense savoring and much lingering. Take as an example Shafer's succinct summary of the three branches of the art of conjecture of d'Alembert, which essentially lays out the contours and interplay among (precise) probability, statistics, and imprecise probability. The first branch enters the game of conjecture by manipulating theoretically precisely specified quantities and models, such as the proposition of equally likely outcomes. This is essentially the game of precise probability, deducing properties and consequences of a theoretical construct.

The second branch plays the same game empirically, by focusing on assessing and approximating chances and risks from data. This captures the essences of much of the current statistical practice, when such empirical assessments are guided by the rules of precise probability. Principled statistical practices fully recognize the multiple uncertainties in the empirical assessments, and hence have build-in risk assessments for estimating the part of the uncertainties that can be reasonably gauged empirically. For parts that cannot be empirically assessed internally, sensitivity studies have been the primary tool, precisely because by posting specific alternative scenarios, we can traverse within the first two branches, and hence remain in our comfort zone.

Shafer's summary makes it clear that the third branch is where our need to pay far more attention than we currently do, because it is the branch that gives

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

deep trouble–and much life–to our beloved subjects of precise probability and statistics. This branch covers the vast majorities of inquires where precise probabilistic descriptions, whether theoretical or empirical, are inherently incomplete or impossible. In our own applied work (e.g., in astronomy or psychiatry), there has been no exception that when we ask a subject expert to provide a prior, the most precise answer would be of the kind "I'm quite sure that α is between 1 and 2." Any further inquiry about how α is distributed on $[1,2]$ would be met with either a puzzling face or an answer few of us like: "I have no idea."

Such "no-idea" answers have motivated many to work harder throughout history, from seeking deeper theoretical understanding to better empirical learning. Nevertheless, comparing the number of inquires where "no-idea" is the real obstacle, to cases where it is a mild inconvenience, is a bit like comparing irrational numbers to rational ones: the latter may appear to be everywhere, but they are of no detectable measure compared to the former. As a result, the vast majority of time we are forced to make up assumptions, such as α is distributed uniformly on $[1,2]$, for the sole purpose of applying available theories or methods. Or as Shafer puts it, despite the effort made through out the history to move bits of the third branch into the first two, "the third still seems very large."

Instead of cutting foot to fit shoes, the framework *imprecise probability* (IP) suggests a less painful paradigm: expanding the shoes to fit the foot. This metaphor has another leg to stand on because the imprecise shoes are no less functional than the precise ones. As Augustin and Schollmeyer emphasized, IP should have been better named as "set-valued probability." But sound statistical inference should also be (at least) set-valued in order to reflect uncertainty, as classic inference paradigms have delivered via confidence intervals, Bayesian creditable sets, and so on. In that sense, the set-valued output of IP models is no less familiar a mathematical form than that from precise probability models, albeit carrying a different interpretation of "uncertainty". It is therefore very natural for us to ask why can't we go from set-valued input to set-valued output directly, instead of constructing a fake id—so to speak—in order to gain access to the narrow tunnel of numerically valued probability?

2. THE TWO CONCERNS THAT MOTIVATED OUR WORK

To answer this question, and in light of the rich scope of topics that the discussions together encompass, we would like to elaborate further our view on the role of imprecise probabilities and their accompanying updating rules. Our view is not in fundamental disagreement with that of our discussants, nevertheless it is focused distinctively on the modern practice of statistical inference. Our concern is centered around two arguments in near contradiction to one another.

We surmise nearly all statisticians take it for granted that probability is the language of uncertainty. By probability, the majority of statisticians specifically mean the kind of probability that obeys the Kolmogorov axioms dictating countable additivity. Bayesians, Frequentists, as well as those who entertain fiducial, structural and functional inference, all operate within a framework that guides the expression of uncertainty that relates observable information to unknown quantities of interest, and in this sense, *update* their knowledge in

light of what's been learned.

Probability is the magical lasso with which statisticians tame unruly variabilities. We use probability to gloss over variabilities that we would rather not delve deeper to understand. In nearly every statistical model, we invoke some notion of independence or exchangeability assumption motivated by our ignorance on any relational information, often rightly so. In fact, we use this "ignorance" to our advantage: the theory of randomization is founded on variabilities, artificially induced in a way that make us conclude that we are better off not to delve deeper. The essence of every statistical analysis relies on the judicious reduction of the unknown and the unknowable into a known or knowable probability. One of us discussed (Meng, 2014; K. Liu & Meng, 2016; Li & Meng, 2021) the multi-resolution nature of inference, which gives meaning to the word "judicious" in this context via the choice of the resolution level. Reduction is great - until when it has gone too far to the point of absurdity. This is where imprecise probabilities comes in, and the primary motivation for our work in the first place.

In general, we are concerned with the change of the "model" $P(A)$ to $P_B(A)$, after B has been learned. We borrow here Hausdorff's notation for relative probability, introduced by Shafer. The Bayes Rule, namely the assertion that upon observing event B, the agent shall replace her prior belief $P(A)$ about A with the conditional probability $P(A \mid B)$, has been justified as the rule to instill such a change, even in the context that such learning is to happen dynamically over time. Calling $Cr_{X,T}(H)$ the degree of credence of agent X towards assertion H at time T, and $Cr_{X,T}(H \mid E)$ the conditional credence if X ascertained that proposition E holds, Carnap (1962) maintained that if E is the observational data received by X between times t_1 and t_2, the rational and coherent agent X must transform their credence at time t_1 to time t_2 as $Cr_{X,t_2}(H) = Cr_{X,t_1}(H \mid E)$. Teller (1973)'s dynamic Dutch book argument attempted to compel the same conclusion. He argued that if an agent engages in a mixture of regular and called-off bets at prices that differ from their assessed marginal and conditional probabilities of the uncertain outcome, they would be made a sure loser by the exploitative (and know-it-all, we must add) bookie.

These fine arguments that are applicable only within a limited and highly idealized scope. As compelling as they may be, the proposal to dynamically update one's credence via Bayesian conditionalization require that the agent knows the "full road map" ahead of them. The proposal does not address the complexity of uncertainty reasoning that modern statisticians and quantitative researchers at large typically faced with, that is, when we only have a partial map or no map at all. This is why, first and foremost, we celebrate the potential of IP tools in resolving this matter. We are therefore happy to see the support and enhancement from the discussants from multiple angles. Augustin and Schollmeyer provide us with a succinct overview of the use of credal sets in statistical inference, and their envelope representation is particularly appealing because it is rooted in distributional families, a concept familiar to statisticians. Liu and Martin, here and more generally in their work on inferential models (IM; Martin & Liu, 2015), argue that to achieve the validity as they defined, we must resort to IP models. Wheeler opened our eyes further by introducing us to a world with constructs that are even more general and more primitive to

credal sets.

Nevertheless, to invoke IP quantification resolve only one aspect of what we consider as an overly aggressive reduction to uncertainty reasoning that requires precise probabilities. In some–and one may argue, most–situations, the statistician regardless of persuasion must contemplate how to update knowledge, not in terms of uncertainty but rather *in the presence* of uncertainty. By *in the presence* we mean that the analyst is not certain of what model structure they are to construct in the first place, or that they do not have an idea how isolated pieces of information, such as individual observations, interact with each other. In the terminology of Liu and Martin, it is the "association equation" that is ill-defined, or that we know certain margins of the association equation, but not how they behave together jointly. This is why the focus of our article is not on IP description itself, but rather conditioning (and by extension, combining) rules for IP. There, what is our equivalent go-to assumption, like exchangeability in the precise world?

The examples presented in our article were chosen for their simple nature. Thus, when disagreement arise among the rules in question, not only is the effect stark, but also the reader may appeal to their own intuitive judgment as to which answers are more sensible than others. The IM treatment offered by Liu and Martin work well in these situations, because it is *apriori* known to the modeler what kind of inferential conclusion is more desirable. Their model building process, including the specification of the association and the choice of the predictive random set, reflects these convictions of the modeler and produces— without surprise—results that are both desirable and intuitive. Pedagogical examples can only go so far, however. The leap from simple examples to reality deprives the modeler the luxury of intuition, and the decision on how to update becomes nontrivial when it is no longer preceded by the answer. Just like other modes of IP-based frameworks of inference, the IM framework is faced with a nontrivial choice of rules when it comes to combining marginal models (Martin & Syring, 2019). A fundamental question in all these situations is whether this choice shall be guided by so-called desirable properties that pertain to the resulting answer, knowing that our notion of what properties are desirable is riddled with inaccurate assumptions, which IP models intend to address in the first place.

3. PARTIAL ORDERING: IS IT A FEATURE OR A BUG?

Mathematically, the multiplicity of rules is a result of using sets to capture the low-resolution nature of our data or information (Gong & Meng, 2021). Sets only obey partial ordering: a set A can be neither larger nor smaller than another set B. Or in terms of evidence, knowing M is in A may imply neither M is in B nor it is not. But in life, ambiguity is the rule, not the exception. Just as ambiguities in life typically lead to multiple scenarios or considerations, partial ordering permits multiple ways to revise our probabilistic assessments when we need to take into account additional considerations, whether for updating, focusing, or any other reasons that require a reassessment. We therefore argue partial ordering necessitated by treating sets as the fundamental building blocks for probability specifications is rather a feature, not a bug, for dealing with imprecise data, information, or other forms of inconclusive evidence.

Perhaps because of our initial (immature) desire of a single rule of such reassessments for the third branch of art of conjecture, just as Bayes rule does for the first two branches, our article focused mostly on studying and comparing individual rules, instead of seeking deeper unification. We are therefore particularly grateful to Augustin and Schollmeyer for giving us a healthy dosage of depression, as their elegant envelope representation is the unification we missed when we attempt to discuss the "optimism" of Dempster's rule, the "pessimism" of the Geometric rule, and the "conservatism" of the generalized Bayes rule, in the three-prisoner example. Their envelop representation makes it crystal clear that (1) ideological differences are inherently embedded into the rules, hence are omnipresent; and (2) behavioral differences among the rules are driven by their underlying ideology.

The significance of the envelope representation is that it puts all three rules as seeking extreme probabilities in the space of a family of distributions, subject to different (further) constraints. The generalized Bayes rule assumes no further constraint, and hence the resulting set-valued probability $[\underline{P}, \overline{P}]$ is the widest possible, hence the most conservative when we take the width of the interval as a measure of the sharpness of our probabilistic assessment or lack thereof. To better understand the optimism of Dempster's rule and the pessimism of the Geometric rule, especially for readers who are yet comfortable with the language of IP, it helps to consider the case of belief function, corresponding to Choquet capacity of order infinity, where we can map a set of probabilities to an ordinary probability of sets (see e.g., Gong & Meng, 2021).

Specifically, under the precise probability formulation, when we need to reassess a probability by moving from its original state space Ω to a subset $S \subset \Omega$, we will permit and only permit any $\omega \in S$, no less or more. In contrast, in the setting of belief function, "moving from Ω to B" can have multiple interpretations due to the ambiguity reflected by the partial ordering. We can take most generous route by permitting any (non-zero mass) set $A \in \Omega$ that is not ruled out by B, that is, any $A \cap B \neq \emptyset$. This is the route that Dempster's rule takes, and its optimism should be quite evident, since $A \cap B \neq \emptyset$ permits (far) more states ω than $A = B$ would. In contrast, the Geometric rule takes the most restrictive route by only permitting any $A \subseteq B$, that is, a state ω (and its parental set A) is permitted only if it is in B, hence the most pessimistic–or putting it more positively–the most cautious route. As we summarize in our article, these ideological preferences can and often lead to very different results, making the judicious choice of rules a necessary part of the game of inference. Of course, one can pretend to not make a choice by seeking extremes over the rules, but that merely means that one has decided to adopt the generalized Bayes rule, as the envelope representation implies. This, in our view, brings another kind of trouble, as we will discuss in Section 5.

4. ARE WE TOO PESSIMISTIC?

Wheeler and Liu-Martin cast their discussions at very different levels and from very different perspectives. Wheeler's supplied rich background from the IP literature accompanied by a logician's rigor and insight. Liu and Martin took an operational perspective with an utilitarian flavor. They, however, reached essentially the same conclusion, that our article projects a sense of pessimism by

(overly) emphasizing "intrinsic contradictions" with the IP paradigm. Wheeler pointed out that we missed the entire contemporary theory of *lower previsions*, which includes lower probabilities as a special case, and where "coherence preservation under inference is inviolable." Liu and Martin criticized us for not imposing criteria of reliability, which could cure or at least reduce our unsettling feeling.

Wheeler was entirely correct that we missed the theory of lower previsions, because we simply did not know. We are frustrated by our ignorance, and the long learning curve. This is a reflection of Shafer's observation that "the theory of imprecise probability has flourished for several decades, but largely outside of statistics journals." We therefore particularly appreciate the editors of *Statistical Science*, especially executive editors Cunhui Zhang and Sonia Petrone, for seeing the value of this topic and for organizing this discussion, which also provides us with a great learning opportunity.

Liu and Martin were also correct that we did not explicitly impose any reliability criteria. We of course agree that whatever rules one puts forward, there must be some rationales to justify them. In the sentence Liu and Martin quoted, we made it explicit that the rules are "pre-specified", and we consider the choices of criterion a part of the pre-specification. But we did not elaborate on any additional choices of the criteria because of the nature of the topic we present. To us, precise probability is the grammar for statistical inference under the highest-resolution specifications, that is, when we can–or pretend we can–postulate probability specifications on all individual elements in however complicated or high-dimensional joint state spaces, for all quantities involved in our inference. The Bayes theorem is a consequence of the precise probability as we all taught. Adopting Bayes theorem as a rule, or rather adopting the Bayes rule as suggested by the Bayes theorem, can be viewed as a reliable, criteria-driven exercise (e.g., by imposing the coherence requirement). But as Shafer correctly pointed out, the distinction between Bayes rule and Bayes theorem has been essentially ignored in the statistical literature (and we certainly accept Shafer's criticism for our own "confusing" mix of the two). We surmise this was largely due to acceptance of precise probability as a *reliable* grammar for statistical inference, and hence any rules set or implied by it would be accepted without the need for further criterion to justify them.

An initial motivation for our work was our desire to learn what is the natural *generalization* of the Bayes rule, as given by Shafer's (1), in the world of imprecise probabilities. The singular form of "generalization" was intended, as for decades, one of us believed (or hoped) that Dempster's rule was *the* natural generalization of Bayes rule, implied by some "Dempster theorem", a consequence of the belief function apparatus. The dream was broken when the other of us actually studied the issue (not just dreamt about it), and what we presented was a part of that broken dream.

We were therefore hoping that we were wrong, and that our "pessimism" was a result of our ignorance, that is, we had not looked hard enough. Consequently, we were excited initially by Wheeler's emphasis that what we explored only reflected what was known before either of us was born. Whereas we definitely want to and will study the more contemporary theory of lower previsions, a more careful reading of Wheeler's discussion reignited our un-

settling feeling because the lower previsions only preserves the coherence, and Wheeler's conclusion is that the generalized Bayes rule is still his preference. That is, our pessimism is not a reflection of our ideology, but fundamental to the marriage of coherence with IP, as we explain below.

5. SHOULD WE ALSO AVOID FATAL ATTRACTION?

If coherence is the only desirable criterion, we would feel comfortable to settle with the generalized Bayes rule, especially as there is no other (common) rule which possesses that property. However, generalized Bayes rule suffers from a flaw that in our view is no less serious or fatal than being incoherent, that it, it cannot get itself out of the vacuous state of knowledge, regardless the amount of data or information one accumulates. In other words, the vacuous state is a fatal attraction state of the generalized Bayes rule. If we insist on having rules that avoid this fatal attraction, which we see no practical or logical justification, then generalized Bayes rule would be out of the window. We want to emphasize that the "vacuous state" is not a straw man. Much of "objective Bayes" or fiducial inference hope to conduct distributional inference without imposing any prior knowledge, that is, to start with the vacuous state. Any updating rule that has the vacuous state as its fatal attraction clearly will be eliminated from the start.

Similarly, the fact that Liu and Martin's "validity" requirement can lead to vacuous state as the only solution (e.g., see Section 4.3 of Liu and Martin) raises the question of the general desirability or even the validity of this "validity" requirement. Indeed, Liu-Martin's validity requirement is fundamentally a frequentist calibration construct, like unbiasedness for testing. Hence it inherits the known defects of their classic counterparts (e.g., controlling Type I error), such as a lack of relevance, or making the wrong trade-off by assigning higher confidence to harder problems, as discussed in K. Liu & Meng (2016).

We raise these points not to suggest that we have better solutions, but to reaffirm our message that judicious judgment and choices are inevitable. In the grand scheme of things, this emphasis itself is vacuous, since inference is not possible without making any assumption, and any assumption is a judgmental call, judicious or not. The apparent dominating emphasis on coherence in the IP literature suggests that itself is a choice, very judicious indeed. But that does not suggest that it is necessarily *coherent* with other considerations, such as avoiding fatal attraction of the vacuous state. If generalized Bayes rule is the only sensible one by the coherence requirement, then indeed we have to accept the intrinsic contradiction between being coherent and avoiding fatal attraction.

6. THE DEMAND (AND SUPPLY) FOR JUDICIOUS JUDGMENT

Give judicious judgment is inevitable, the key question then is how do we make them? Even in the precise probability situation, often is the situation that the analyst does not quite know what model to specify, except that their partial and meta-knowledge makes them realize that the Bayesian recipe of conditionalization may *not* be the right thing to do under the model they are forced to come up with. In these practical situations, the analyst is not so much motivated by coherence – for the price may well be too high to pay, as Wheeler pointed

out, but rather regard their experience and expertise as meta-information to guide their constructions of ad hoc fixes to deal with the lack of information.

An example of this is the literature on modularized Bayesian inference and cut distributions (F. Liu et al., 2009; Lunn et al., 2009; Plummer, 2015; Jacob et al., 2017), inspired by Bayesian pharmacokinetics and pharmacodynamics (PKPD) models. The analyst has information that a certain margin of the joint model may involve poor quality data or information, and would like to "cut off" the contribution of this margin into other parts of the model for which the analyst has scientific interest. In theory, if we know how to quantify data or information quality, we can incorporate such quantification properly in our probabilistic model. Then following the Bayesian recipe, such as conditioning, would lead to a sensible inference that properly weights various pieces of information by their quality index. However, other than for linear estimates and estimators (Meng, 2018), quantifying deterioration in quality due to non-sampling mechanisms is currently out of reach in theory and in practice.

A common practical approach is then to attach zero weight to the problematic aspects of the data or model, that is, to "cut them off." This is often a better strategy than keeping them, because zero weight is likely a better approximation to the (unknown) optimal weights than blindly pretending that all parts of the data or model should be given the standard treatments, e.g., equal weighting. Evidently, such "updating" procedures through cut distributions do not conform to Bayesian conditionalization. However, they tend to yield better results in practice, because they are better approximations to the optimal but inoperable Bayesian approach under the fully correctly specified model, than mechanically applying the Bayesian rule to the mis-specified model.

We therefore thank Augustin and Schollmeyer again for their proposal of soft revision, as a customizable updating rule that bridges the pessimist Geometric rule and the optimist Dempster's rule, to which they drew a connection to the maximum Bayes factor approach of Good (1967) and an analogy with empirical Bayes. We were reminded of Lindley's declaration that "there is no one less Bayesian than an empirical Bayesian" (Lindley, 1969, in discussion of Copas (1969)). As much as the proposed soft revision ventures outside the realm of coherent Bayesianism, it is nevertheless a useful and welcome addition to the toolbox of the practical statistician, one that could help us avoid the fatal attraction.

As we call for the use of imprecise probabilities, which invariably relies on some kind of new guiding principle to account for new information, call it a rule, protocol, or otherwise, there is risk in harming the operationalizability of the statistical inference framework. We view this is yet another instance of the omnipresent no-free lunch principle. Indeed, there is a price to pay even for every precise generalization of the ordinary probability calculation. Good (1966) advocated for *probabilities of higher types* as a candidate measure of non-measurable events. He remarked, however, that probabilities of higher types are expressible only in terms of inequalities that are fuzzy in nature, and quickly lose practical importance the higher in type they go. Similarly, if we let go the notion of a relatively well-defined protocol, and possibly other aspects of routine practice of model building, it would not be long before a necessary level of operationalizability is lost. When that happens, any theory without

computational feasibility can deter practically minded users even if they are sympathetic to the idea.

As widely endorsed as Bayesian analysis among applied statisticians, the computational challenge was once insurmountable. If it wasn't for the MCMC revolution in the 1990s, Bayesian statistical methods would not have taken off, at least not this rapidly. Customizable computational apparatuses, such as WinBUGS and Stan, made Bayesian computation on large scale datasets possible. By way of contrast, computation for IP models in statistical inference is still in its early development. The SIPTA (Society for Imprecise Probability: Theories and Applications) community has seen recent advances on the use of MCMC to estimate lower expectations (Fetz, 2019; Decadt et al., 2019). The statistical literature is starting to catch up in that regard. The recent work of Jacob et al. (2021) developed the first workable sampler for the random convex polytope proposed fifty years prior (Dempster, 1966, 1972), characterizing the Dempster-Shafer inference for categorical data.

The motto to the SIPTA community is that "there are more uncertainties than probabilities". As statisticians, we are eager to see it reflected in the practice of statistical inference. By conducting imprecise probability inference, we can be precise about what we do not know, an hence deliver more replicable results because we avoid making up assumptions forced upon by the limitations of the precise probability framework. We therefore invite anyone who cares about scientific replicability to look into what the world of IP can offer.

REFERENCES

Carnap, R. (1962). The aim of inductive logic. In E. Nagel, P. Suppes, & A. Tarski (Eds.), *Logic, methodology and philosophy of science* (pp. 303–318). Stanford, CA: Stanford University Press.

Copas, J. (1969). Compound decisions and empirical bayes. *Journal of the Royal Statistical Society: Series B (Methodological)*, *31*(3), 397–417.

Decadt, A., De Cooman, G., & De Bock, J. (2019). Monte carlo estimation for imprecise probabilities: Basic properties. In *International symposium on imprecise probabilities: Theories and applications* (pp. 135–144).

Dempster, A. P. (1966). New methods for reasoning towards posterior distributions based on sample data. *The Annals of Mathematical Statistics*, *37*(2), 355–374.

Dempster, A. P. (1972). A class of random convex polytopes. *The Annals of Mathematical Statistics*, 260–272.

Fetz, T. (2019). Improving the convergence of iterative importance sampling for computing upper and lower expectations. In *International symposium on imprecise probabilities: Theories and applications* (pp. 185–193).

Gong, R., & Meng, X.-L. (2021). Probabilistic underpinning of imprecise probability for statistical learning with low-resolution information. *Technical Report*.

Good, I. J. (1966). Symposium on current views of subjective probability: Subjective probability as the measure of a non-measurable set. In *Studies in logic and the foundations of mathematics* (Vol. 44, pp. 319–329). Elsevier.

Good, I. J. (1967). A bayesian significance test for multinomial distributions. *Journal of the Royal Statistical Society: Series B (Methodological)*, *29*(3), 399–418.

Jacob, P. E., Gong, R., Edlefsen, P. T., & Dempster, A. P. (2021). A Gibbs sampler for a class of random convex polytopes. *Journal of the American Statistical Association (forthcoming, with discussion)*.

Jacob, P. E., Murray, L. M., Holmes, C. C., & Robert, C. P. (2017). Better together? statistical learning in models made of modules. *arXiv preprint arXiv:1708.08719*.

Li, X., & Meng, X.-L. (2021). A multi-resolution theory for approximating infinite-p-zero-n: Transitional inference, individualized predictions, and a world without bias-variance tradeoff. *Journal of the American Statistical Association*, *doi: 10.1080/01621459.2020.1844210*.

Lindley, D. (1969). Discussion of compound decisions and empirical bayes, jb copas. *Journal of the Royal Statistical Society, Series B*, *31*, 419–421.

Liu, F., Bayarri, M., Berger, J., et al. (2009). Modularization in bayesian analysis, with emphasis on analysis of computer models. *Bayesian Analysis*, *4*(1), 119–150.

Liu, K., & Meng, X.-L. (2016). There is individualized treatment. Why not individualized inference? *The Annual Review of Statistics and Its Applications*, *3*, 79-111.

Lunn, D., Best, N., Spiegelhalter, D., Graham, G., & Neuenschwander, B. (2009). Combining mcmc with 'sequential' pkpd modelling. *Journal of Pharmacokinetics and Pharmacodynamics*, *36*(1), 19.

Martin, R., & Liu, C. (2015). *Inferential Models: reasoning with uncertainty*. CRC Press, Boca Raton, FL.

Martin, R., & Syring, N. (2019). Validity-preservation properties of rules for combining inferential models. In *International symposium on imprecise probabilities: Theories and applications* (pp. 286–294).

Meng, X.-L. (2014). A trio of inference problems that could win you a nobel prize in statistics (if you help fund it). *Past, Present, and Future of Statistical Science*, 537–562.

Meng, X.-L. (2018). Statistical paradises and paradoxes in big data (i): Law of large populations, big data paradox, and the 2016 us presidential election. *Annals of Applied Statistics*, *12*(2), 685–726.

Plummer, M. (2015). Cuts in bayesian graphical models. *Statistics and Computing*, *25*(1), 37–43.

Teller, P. (1973). Conditionalization and observation. *Synthese*, *26*(2), 218–258.