1 Introduction

The geometry of reason refers to a view of epistemic utility in which the underlying topology for credence functions (which may be subjective probability distributions) on a finite number of events is a metric space. The set of non-negative credences that an agent assigns to the outcome of a die roll, for example, is isomorphic to $\mathbb{R}^6_{\geq 0}$. If the agent fulfills the requirements of probabilism, the isomorphism is to the more narrow set $\mathbb{S}^5 \subset \mathbb{R}^6$, for which the coordinates sum to 1.

For the remainder of this paper I will assume probabilism and an isomorphism between probability distributions P on an outcome space Ω with $|\Omega| = n$ and points $p \in \mathbb{S}^{n-1} \subset \mathbb{R}^n$ having coordinates $p_i = P(\omega_i), i = 1, \ldots, n$ and $\omega_i \in \Omega$. Since the isomorphism is to a metric space, there is a distance relation between credence functions which can be used to formulate axioms relating credences to epistemic utility and to justify or to criticize contentious positions such as Bayesian conditionalization, the principle of indifference, other forms of conditioning, or probabilism itself (see especially works cited below by James Joyce; Pettigrew and Leitgeb; David Wallace and Hilary Greaves). For information theory, as opposed to the geometry of reason, the underlying topology for credence functions is not a metric space (see figures 1 and 2 for illustration).

Leitgeb and Pettigrew show, given the geometry of reason, Joyce's 'Norm of Gradational Accuracy' (see Joyce, 1998, 579) and other axioms inspired by the epistemic utility approach, that in order to avoid epistemic dilemmas we must commit ourselves to a Brier score measure of inaccuracy and subsequently to probabilism and standard conditioning. The Brier score is the mean squared error of a probabilistic forecast. For example, if we look at 100 days for which the forecast was 30% rain and the incidence of rain was 32 days, then the Brier score is

$$\frac{1}{N}\sum_{t=1}^{N}(f_t - o_t) = \frac{1}{100}\left(32 \cdot (0.3 - 1) + 68 \cdot (0.3 - 0)\right) = 0.218 \qquad (1)$$

0 is a perfect match between forecast and reality in the sense that the forecaster anticipates every instance of rain with a 100% forecast and every instance of no rain with a 0% forecast.

Jeffrey conditioning (also called probability kinematics) is widely considered to be a commonsense extension of standard conditioning. On Leitgeb and

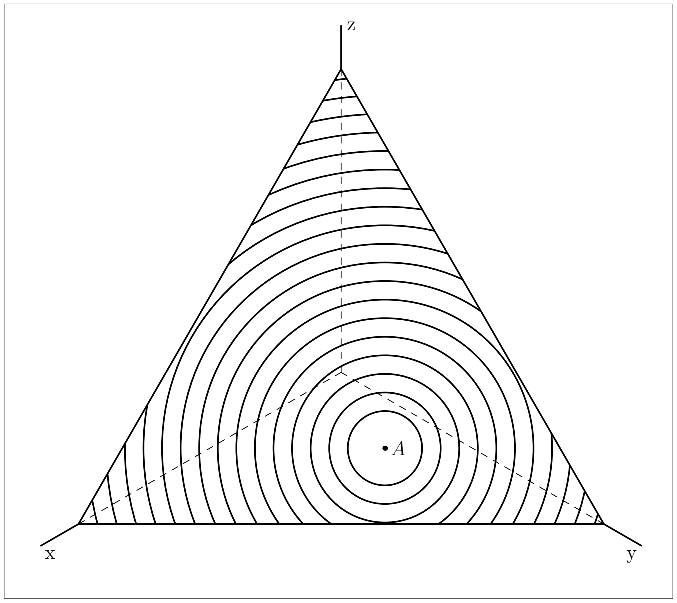

Figure 1: The simplex $\mathbb{S}^2$ in three-dimensional space $\mathbb{R}^3$ with contour lines corresponding to the geometry of reason around point A in equation (2). Points on the same contour line are equidistant from A with respect to the Euclidean metric. Compare the contour lines here to figure 2. Note that this diagram and all the following diagrams are frontal views of the simplex.

Pettigrew's account, using the Brier score, it fails to provide maximal epistemic utility. Another type of conditioning, which I will call LP conditioning, takes the place of Jeffrey conditioning. The failure of Jeffrey conditioning to minimize inaccuracy on the basis of the geometry of reason casts doubt on the geometry of reason by reductio.

2 Epistemic Utility and the Geometry of Reason

2.1 Epistemic Utility for Partial Beliefs

Joyce and Leitgeb/Pettigrew propose various axioms for a measure of gradational accuracy for partial beliefs relying on the geometry of reason, i.e. the idea of geometrical distance between distributions of partial belief expressed in non-negative real numbers. In Joyce, a metric space for probability dis-

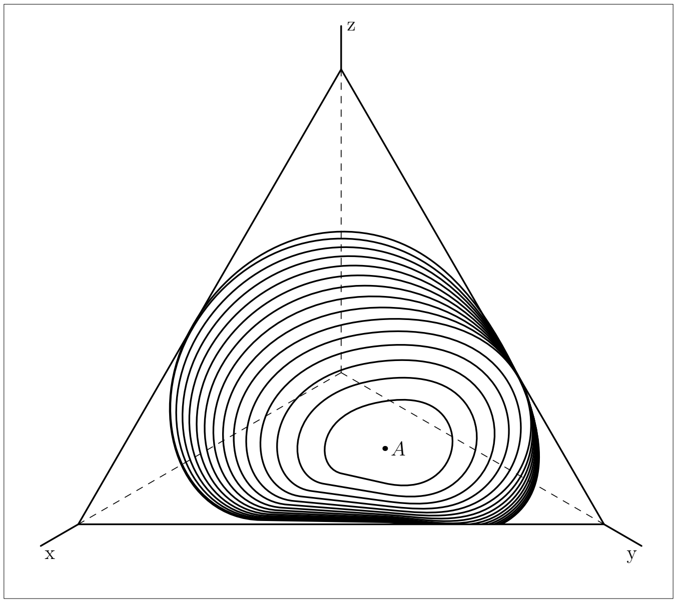

Figure 2: The simplex $\mathbb{S}^2$ with contour lines corresponding to information theory around point A in equation (2). Points on the same contour line are equidistant from A with respect to the Kullback-Leibler divergence. The contrast to figure 1 will become clear in much more detail in the body of the paper. Note that the contour lines of the geometry of reason are insensitive to the boundaries of the simplex, while the contour lines of information theory reflect them. One of the main arguments in this paper is that information theory respects epistemic intuitions we have about asymmetry: proximity to extreme beliefs with very high or very low probability influences the topology that is at the basis of updating.

tributions is adopted without much reflection. The midpoint between two points, for example, which is freely used by Joyce, assumes symmetry between the end points. The asymmetric divergence measure that I propose as an alternative to the Euclidean distance measure has no meaningful concept of a midpoint.

Leitgeb and Pettigrew muse about alternative geometries, especially non-Euclidean ones. They suspect that these would be based on and in the end reducible to Euclidean geometry but they do not entertain the idea that they could drop the requirement of a metric topology altogether (for the use of non-Euclidean geodesics in statistical inference see Amari, 1985). Thomas

Mormann explicitly warns against the assumption that the metrics for a geometry of logic is Euclidean by default, “All too often, we rely on geometric intuitions that are determined by Euclidean prejudices. The geometry of logic, however, does not fit the standard Euclidean metrical framework” (see Mormann, 2005, 433; also Miller, 1984). Mormann concludes in his article “Geometry of Logic and Truth Approximation,”

> Logical structures come along with ready-made geometric structures that can be used for matters of truth approximation. Admittedly, these geometric structures differ from those we are accostumed [sic] with, namely, Euclidean ones. Hence, the geometry of logic is not Euclidean geometry. This result should not come as a big surprise. There is no reason to assume that the conceptual spaces we use for representing our theories and their relations have an [sic] Euclidean structure. On the contrary, this would appear to be an improbable coincidence. (Mormann, 2005, 453)

2.2 Axioms for Epistemic Utility

To give a flavour of how attached the axioms are to the geometry of reason, here are Joyce’s axioms called Weak Convexity and Symmetry, which he uses to justify probabilism.

> **Weak Convexity**: Let $m = (0.5b' + 0.5b'')$ be the midpoint of the line segment between b' and b''. If $I(b', \omega) = I(b'', \omega)$, then it will always be the case that $I(b', \omega) \geq I(m, \omega)$ with identity only if $b' = b''$.

> **Symmetry**: If $I(b', \omega) = I(b'', \omega)$, then for any $\lambda \in [0, 1]$ one has
> $I(\lambda b' + (1 - \lambda)b'', \omega) = I((1 - \lambda)b' + \lambda b''), \omega)$.

Joyce advocates for these axioms in Euclidean terms, using justifications such as “the change in belief involved in going from b' to b'' has the same direction but a doubly greater magnitude than change involved in going from b' to [the midpoint] m” (see Joyce, 1998, 596). In section 5.3, I will show that Weak Convexity holds, and Symmetry does not hold, in ‘information geometry,’ the topology generated by the Kullback-Leibler divergence. The term information geometry is due to Imre Csiszár, who considers the Kullback-Leibler divergence a non-commutative (asymmetric) analogue of squared Euclidean distance and derives several results that are intuitive information geometric counterparts of standard results in Euclidean geometry (see chapter 3 of Csiszár and Shields, 2004).

2.3 Expectations for Jeffrey-Type Updating Scenarios

Jeffrey conditioning does not fulfill Joyce's Norm of Gradational Accuracy and therefore violates the pursuit of epistemic utility. Leitgeb and Pettigrew provide us with an alternative method of updating for Jeffrey-type updating scenarios, which I will call LP conditioning.

> **Example 1: Sherlock Holmes.** Sherlock Holmes attributes the following probabilities to the propositions E_i that k_i is the culprit in a crime: $P(E_1) = 1/3, P(E_2) = 1/2, P(E_3) = 1/6$, where k_1 is Mr. R., k_2 is Ms. S., and k_3 is Ms. T. Then Holmes finds some evidence which convinces him that $P'(F^*) = 1/2$, where F^* is the proposition that the culprit is male and P is relatively prior to P'. What should be Holmes' updated probability that Ms. S. is the culprit?

I will look at the recommendations of Jeffrey conditioning and LP conditioning for example 1 in the next section. For now note that LP conditioning violates all of the following plausible expectations in **List A** for an amujus, an 'alternative method of updating for Jeffrey-type updating scenarios.' This is **List A**:

- CONTINUITY An amujus ought to be continuous with standard conditioning as a limiting case.
- REGULARITY An amujus ought not to assign a posterior probability of 0 to an event which has a positive prior probability and about which the intervening evidence says nothing except that a strictly weaker event has a positive posterior probability.
- LEVINSTEIN An amujus ought not to give "extremely unattractive" results in a Levinstein scenario (see Levinstein, 2012, which not only articulates this failed expectation for LP conditioning, but also the previous two).
- INVARIANCE An amujus ought to be partition invariant.

- EXPANSIBILITY An amujus ought to be insensitive to an expansion of the event space by zero-probability events.

- HORIZON An amujus ought to exhibit the horizon effect which makes probability distributions which are nearer to extreme probability distributions appear to be closer to each other than they really are.

Jeffrey conditioning and LP conditioning are both an amujus based on a concept of quantitative difference between probability distributions measured as a function on the isomorphic manifold (in our case, an $n-1$-dimensional simplex). Evidence appears in the form of a constraint on acceptable probability distributions and the closest acceptable probability to the orginal (relatively prior) probability distribution is chosen as its successor. Here is **List B**, a list of reasonable expectations one may have toward this concept of quantitative difference (we call it a distance function for the geometry of reason and a divergence for information theory). Let $d(p,q)$ express this concept mathematically.

- TRIANGULARITY The concept obeys the triangle inequality. If there is an intermediate probability distribution, it will not make the difference smaller: $d(p,r) \leq d(p,q) + d(q,r)$. Buying a pair of shoes is not going to be more expensive than buying the two shoes individually.

- COLLINEAR HORIZON This expecation is just a more technical restatement of the HORIZON expectation in the previous list. If p, p', q, q' are collinear with the centre of the simplex m (whose coordinates are $m_i = 1/n$ for all i) and an arbitrary but fixed boundary point $\xi \in \partial\mathbb{S}^{n-1}$ and p, p', q, q' are all between m and ξ with $\|p'-p\| = \|q'-q\|$ where p is strictly closest to m, then $|d(p,p')| < |d(q,q')|$. For an illustration of this expectation see figure 3.

- TRANSITIVITY OF ASYMMETRY An ordered pair (p,q) of simplex points associated with probability distributions is asymmetrically negative, positive, or balanced, so either $d(p,q)-d(q,p) < 0$ or $d(p,q)-d(q,p) > 0$ or $d(p,q)-d(q,p) = 0$. If (p,q) and (q,r) are asymmetrically positive, (p,r) ought not to be asymmetrically negative. Think of a bicycle route map with different locations at varying altitudes. If it takes 20 minutes to get from A to B but only 15 minutes to get from B to A then (A,B) is asymmetrically positive. If (A,B) and (B,C) are asymmetrically positive, then (A,C) ought not to be asymmetrically negative.

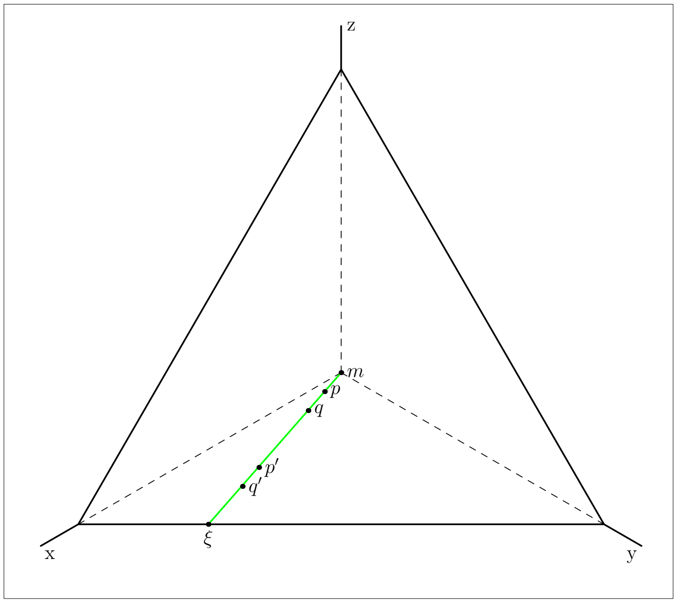

Figure 3: An illustrations of conditions (i)–(iii) for COLLINEAR HORIZON in **List B**. p, p' and q, q' must be equidistant and collinear with m and ξ. If q, q' is more peripheral than p, p', then COLLINEAR HORIZON requires that $|d(p, p')| < |d(q, q')|$.

While the Kullback-Leibler divergence of information theory fulfills all the expectations of **List A**, save HORIZON, it fails all the expectations in **List B**. Obversely, the Euclidean distance of the geometry of reason fulfills all the expectations of **List B**, save COLLINEAR HORIZON, and fails all the expectations in **List A**. Information theory has its own axiomatic approach to justifying probabilism and standard conditioning (see Shore and Johnson, 1980). Information theory provides a justification for Jeffrey conditioning and generalizes it (see Lukits, 2015). All of these virtues stand in contrast to the violations of the expectations in **List B**. The rest of this paper fills in the details of these violations both for the geometry of reason and information theory, with the conclusion that the case for the geometry of reason is hopeless while the case for information theory is now a major challenge for future research projects.

3 Geometry of Reason versus Information Theory

Here is a simple example corresponding to example 1 where the distance of geometry and the divergence of information theory differ. With this difference in mind, I will show how LP conditioning fails the expectations outlined in **List A**. Consider the following three points in three-dimensional space:

$$a = \left(\frac{1}{3}, \frac{1}{2}, \frac{1}{6}\right) \qquad b = \left(\frac{1}{2}, \frac{3}{8}, \frac{1}{8}\right) \qquad c = \left(\frac{1}{2}, \frac{5}{12}, \frac{1}{12}\right) \tag{2}$$

All three are elements of the simplex $\mathbb{S}^2$: their coordinates add up to 1. Thus they represent probability distributions A, B, C over a partition of the event space into three events. Now call $D_{\text{KL}}(B, A)$ the Kullback-Leibler divergence of B from A defined as follows, where a_i are the Cartesian coordinates of a (the base of the logarithm is not important, in order to facilitate easy differentiation I will use the natural logarithm):

$$D_{\text{KL}}(B, A) = \sum_{i=1}^{3} b_i \log \frac{b_i}{a_i}. \tag{3}$$

Note that the Kullback-Leibler divergence, irrespective of dimension, is always positive as a consequence of Gibbs' inequality (see MacKay, 2003, sections 2.6 and 2.7).

The Euclidean distance is defined as follows:

$$\|B - A\| = \sqrt{\sum_{i=1}^{n} (b_i - a_i)^2}. \tag{4}$$

What is remarkable about the three points in (2) is that

$$\|C - A\| \approx 0.204 < \|B - A\| \approx 0.212 \tag{5}$$

and

$$D_{\mathrm{KL}}(B, A) \approx 0.0589 < D_{\mathrm{KL}}(C, A) \approx 0.069. \tag{6}$$

The Kullback-Leibler divergence and Euclidean distance give different recommendations with respect to proximity.

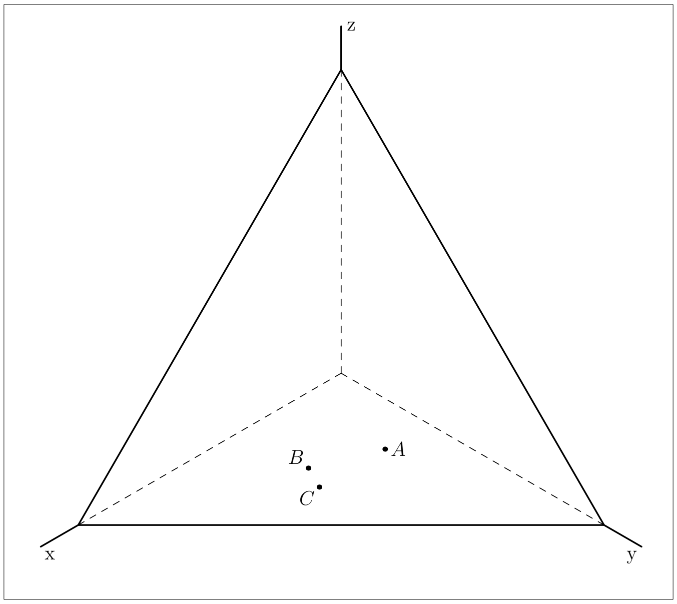

Figure 4: The simplex $\mathbb{S}^2$ in three-dimensional space $\mathbb{R}^3$ with points a, b, c as in equation (2) representing probability distributions A, B, C. Note that geometrically speaking C is closer to A than B is. Using the Kullback-Leibler divergence, however, B is closer to A than C is.

If A corresponds to my prior and my evidence is such that I must change the first coordinate to $1/2$ (as in example 1) and nothing stronger, then information theory via the Kullback-Leibler divergence recommends the posterior corresponding to B; and the geometry of reason as expounded in Leitgeb and Pettigrew recommends the posterior corresponding to C. There are several things going on here that need some explanation.

3.1 LP conditioning and Jeffrey Conditioning

I want to outline how Leitgeb and Pettigrew arrive at posterior probability distributions in Jeffrey-type updating scenarios. I will call their method LP conditioning.

> **Example 2: Abstract Holmes.** Consider a possibility space $W = E_1 \cup E_2 \cup E_3$ (the E_i are sets of states which are pairwise disjoint and whose union is W) and a partition $\mathcal{F}$ of W such that $\mathcal{F} = \{F^*, F^{**}\} = \{E_1, E_2 \cup E_3\}$.

Let P be the prior probability function on W and P' the posterior. I will keep the notation informal to make this simple, not mathematically precise. Jeffrey-type updating scenarios give us new information on the posterior probabilities of partitions such as $\mathcal{F}$. In example 2, let

$$\begin{array}{rcl} P(E_1) & = & 1/3 \\ P(E_2) & = & 1/2 \\ P(E_3) & = & 1/6 \end{array} \tag{7}$$

and the new evidence constrain P' such that $P'(F^*) = 1/2 = P'(F^{**})$.

Jeffrey conditioning works on the following intuition, which elsewhere I have called Jeffrey's updating principle JUP (see also Wagner, 2002). The posterior probabilities conditional on the partition elements equal the prior probabilities conditional on the partition elements since we have no information in the evidence that they should have changed. Hence,

$$\begin{aligned} P'_{\text{JC}}(E_i) \qquad & = P'(E_i|F^*)P'(F^*) + P'(E_i|F^{**})P'(F^{**}) \\ & = P(E_i|F^*)P'(F^*) + P(E_i|F^{**})P'(F^{**}) \end{aligned} \tag{8}$$

Jeffrey conditioning is controversial (for an introduction to Jeffrey conditioning see Jeffrey, 1965; for its statistical and formal properties see Diaconis and Zabell, 1982; for a pragmatic vindication of Jeffrey conditioning see Armendt, 1980, and Skyrms, 1986; for criticism see Howson and Franklin, 1994). Information theory, however, supports Jeffrey conditioning. Leitgeb and Pettigrew show that Jeffrey conditioning does not in general pick out the minimally inaccurate posterior probability distribution. If the geometry of reason as presented in Leitgeb and Pettigrew is sound, this would constitute a powerful criticism of Jeffrey conditioning. Leitgeb and Pettigrew introduce an alternative to Jeffrey conditioning, which I call LP conditioning. It proceeds as follows for example 2 and in general provides the minimally inaccurate posterior probability distribution in Jeffrey-type updating scenarios.

Solve the following two equations for x and y:

$$\begin{array}{rcl} P(E_1)+x & = & P'(F^{*}) \\ P(E_2)+y+P(E_3)+y & = & P'(F^{**}) \end{array} \tag{9}$$

and then set

$$\begin{array}{rcl} P'_{\text{LP}}(E_1) & = & P(E_1)+x \\ P'_{\text{LP}}(E_2) & = & P(E_2)+y \\ P'_{\text{LP}}(E_3) & = & P(E_3)+y \end{array} \tag{10}$$

For the more formal and more general account see Leitgeb and Pettigrew, 2010, 254. The results for example 2 are:

$$\begin{array}{rcl} P'_{\text{LP}}(E_1) & = & 1/2 \\ P'_{\text{LP}}(E_2) & = & 5/12 \\ P'_{\text{LP}}(E_3) & = & 1/12 \end{array} \tag{11}$$

Compare these results to the results of Jeffrey conditioning:

$$\begin{array}{rcl} P'_{\text{JC}}(E_1) & = & 1/2 \\ P'_{\text{JC}}(E_2) & = & 3/8 \\ P'_{\text{JC}}(E_3) & = & 1/8 \end{array} \tag{12}$$

Note that (7), (12), and (11) correspond to A, B, C in (2).

4 Expectations for the Geometry of Reason

This section provides more detail for the expectations in **List A** and shows how LP conditioning violates them.

4.1 Continuity

LP conditioning violates CONTINUITY because standard conditioning gives a different recommendation than a parallel sequence of Jeffrey-type updating scenarios which get arbitrarily close to standard event observation. This is especially troubling considering how important the case for standard conditioning is to Leitgeb and Pettigrew.

To illustrate a CONTINUITY violation, consider the case where Sherlock Holmes reduces his credence that the culprit was male to $\varepsilon_n = 1/n$ for $n = 4, 5, \ldots$. The sequence ε_n is not meant to reflect a case where Sherlock Holmes becomes successively more certain that the culprit was female. It is meant to reflect countably many parallel scenarios which only differ by the degree to which Sherlock Holmes is sure that the culprit was female. These parallel scenarios give rise to a parallel sequence (as opposed to a successive sequence) of updated probabilities $P'_{\text{LP}}(F^{**})$ and another sequence of updated probabilities $P'_{\text{JC}}(F^{**})$ (F^{**} is the proposition that the culprit is female). As $n \to \infty$, both of these sequences go to one.

Straightforward conditionalization on the evidence that ‘the culprit is female’ gives us

$$\begin{array}{rcl} P'_{\text{SC}}(E_1) & = & 0 \\ P'_{\text{SC}}(E_2) & = & 3/4 \\ P'_{\text{SC}}(E_3) & = & 1/4. \end{array} \tag{13}$$

Letting $n \to \infty$ for Jeffrey conditioning yields

$$\begin{array}{rcccl} P'_{\text{JC}}(E_1) & = & 1/n & \to & 0 \\ P'_{\text{JC}}(E_2) & = & 3(n-1)/4n & \to & 3/4 \\ P'_{\text{JC}}(E_3) & = & (n-1)/4n & \to & 1/4, \end{array} \tag{14}$$

whereas letting $n \to \infty$ for LP conditioning yields

$$\begin{array}{lcll} P'_{\text{LP}}(E_1) & = & 1/n & \to 0 \\ P'_{\text{LP}}(E_2) & = & (4n-3)/6n & \to 2/3 \\ P'_{\text{LP}}(E_3) & = & (2n-5)/6n & \to 1/3. \end{array} \tag{15}$$

LP conditioning violates CONTINUITY.

4.2 Regularity

LP conditioning violates REGULARITY because formerly positive probabilities can be reduced to 0 even though the new information in the Jeffrey-type updating scenario makes no such requirements (as is usually the case for standard conditioning). Ironically, Jeffrey-type updating scenarios are meant to be a better reflection of real-life updating because they avoid extreme probabilities.

The violation becomes serious if we are already sympathetic to an information-based account: the amount of information required to turn a non-extreme probability into one that is extreme (0 or 1) is infinite. Whereas the geometry of reason considers extreme probabilities to be easily accessible by non-extreme probabilities under new information (much like a marble rolling off a table or a bowling ball heading for the gutter), information theory envisions extreme probabilities more like an event horizon. The nearer you are to the extreme probabilities, the more information you need to move on. For an observer, the horizon is never reached.

> **Example 3: Regularity Holmes.** Everything is as in example 1, except that Sherlock Holmes becomes confident to a degree of 2/3 that Mr. R is the culprit and updates his relatively prior probability distribution in (7).

Then his posterior probabilities look as follows:

$$\begin{array}{lcl} P'_{\text{JC}}(E_1) & = & 2/3 \\ P'_{\text{JC}}(E_2) & = & 1/4 \\ P'_{\text{JC}}(E_3) & = & 1/12 \end{array} \tag{16}$$

$$\begin{array}{rcl} P'_{\text{LP}}(E_1) & = & 2/3 \\ P'_{\text{LP}}(E_2) & = & 1/3 \\ P'_{\text{LP}}(E_3) & = & 0 \end{array} \tag{17}$$

With LP conditioning, Sherlock Holmes' subjective probability that Ms. T is the culprit in example 3 has been reduced to zero. No finite amount of information could bring Ms. T back into consideration as a culprit in this crime, and Sherlock Holmes should be willing to bet any amount of money against a penny that she is not the culprit—even though his evidence is nothing more than an increase in the probability that Mr. R is the culprit.

LP conditioning violates REGULARITY.

4.3 Levinstein

LP conditioning violates LEVINSTEIN because of "the potentially dramatic effect [LP conditioning] can have on the likelihood ratios between different propositions" (Levinstein, 2012, 419). Consider Benjamin Levinstein's example:

> **Example 4: Levinstein's Ghost.** There is a car behind an opaque door, which you are almost sure is blue but which you know might be red. You are almost certain of materialism, but you admit that there's some minute possibility that ghosts exist. Now the opaque door is opened, and the lighting is fairly good. You are quite surprised at your sensory input: your new credence that the car is red is very high.

Jeffrey conditioning leads to no change in opinion about ghosts. Under LP conditioning, however, seeing the car raises the probability that there are ghosts to an astonishing 47%, given Levinstein's reasonable priors. Levinstein proposes a logarithmic inaccuracy measure as a remedy to avoid violation of LEVINSTEIN. As a special case of applying a Levinstein-type logarithmic inaccuracy measure, information theory does not violate LEVINSTEIN.

4.4 Invariance

LP conditioning violates INVARIANCE because two agents who have identical credences with respect to a partition of the event space may disagree about this partition after LP conditioning, even when the Jeffrey-type updating scenario provides no new information about the more finely grained partitions on which the two agents disagree.

> **Example 5: Jane Marple.** Jane Marple is on the same case as Sherlock Holmes in example 1 and arrives at the same relatively prior probability distribution as Sherlock Holmes (I will call Jane Marple's relatively prior probability distribution Q and her posterior probability distribution Q'). Jane Marple, however, has a more finely grained probability assignment than Sherlock Holmes and distinguishes between the case where Ms. S went to boarding school with her, of which she has a vague memory, and the case where Ms. S did not and the vague memory is only about a fleeting resemblance of Ms. S with another boarding school mate. Whether or not Ms. S went to boarding school with Jane Marple is completely beside the point with respect to the crime, and Jane Marple considers the possibilities equiprobable whether or not Ms. S went to boarding school with her.

Let $E_2 \equiv E_2^* \vee E_2^{**}$, where E_2^* is the proposition that Ms. S is the culprit and she went to boarding school with Jane Marple and E_2^{**} is the proposition that Ms. S is the culprit and she did not go to boarding school with Jane Marple. Then

$$\begin{array}{rcl} Q(E_1) & = & 1/3 \\ Q(E_2^*) & = & 1/4 \\ Q(E_2^{**}) & = & 1/4 \\ Q(E_3) & = & 1/6. \end{array} \tag{18}$$

Now note that while Sherlock Holmes and Jane Marple agree on the relevant facts of the criminal case (who is the culprit?) in their posterior probabilities if they use Jeffrey conditioning,

$$\begin{array}{rcl} P'_{\mathrm{JC}}(E_1) & = & 1/2 \\ P'_{\mathrm{JC}}(E_2) & = & 3/8 \\ P'_{\mathrm{JC}}(E_3) & = & 1/8 \end{array} \tag{19}$$

$$\begin{array}{rcl} Q'_{\mathrm{JC}}(E_1) & = & 1/2 \\ Q'_{\mathrm{JC}}(E_2^*) & = & 3/16 \\ Q'_{\mathrm{JC}}(E_2^{**}) & = & 3/16 \\ Q'_{\mathrm{JC}}(E_3) & = & 1/8 \end{array} \tag{20}$$

they do not agree if they use LP conditioning,

$$\begin{array}{rcl} P'_{\mathrm{LP}}(E_1) & = & 1/2 \\ P'_{\mathrm{LP}}(E_2) & = & 5/12 \\ P'_{\mathrm{LP}}(E_3) & = & 1/12 \end{array} \tag{21}$$

$$\begin{array}{rcl} Q'_{\mathrm{LP}}(E_1) & = & 1/2 \\ Q'_{\mathrm{LP}}(E_2^*) & = & 7/36 \\ Q'_{\mathrm{LP}}(E_2^{**}) & = & 7/36 \\ Q'_{\mathrm{LP}}(E_3) & = & 1/9. \end{array} \tag{22}$$

LP conditioning violates INVARIANCE.

4.5 Expansibility

One particular problem with the lack of invariance for LP conditioning is how zero-probability events should be included in the list of prior probabilities that determines the value of the posterior probabilities. Consider

$$\begin{array}{rcl} P(X_1) & = & 0 \\ P(X_2) & = & 0.3 \\ P(X_3) & = & 0.6 \\ P(X_4) & = & 0.1 \end{array} \tag{23}$$

That $P(X_1) = 0$ may be a consequence of standard conditioning in a previous step. Now the agent learns that $P'(X_3 \vee X_4) = 0.5$. Should the agent update on the list presented in (23) or on the following list:

$$\begin{array}{rcl} P(X_2) & = & 0.3 \\ P(X_3) & = & 0.6 \\ P(X_4) & = & 0.1 \end{array} \tag{24}$$

Whether you update on (23) or (24) makes no difference to Jeffrey conditioning, but due to the lack of invariance it makes a difference to LP conditioning, so the geometry of reason needs to find a principled way to specify the appropriate prior probabilities. The only non-arbitrary way to do this is either to include or to exclude all zero probability events on the list. This strategy, however, sounds ill-advised unless one signs on to a stronger version of REGULARITY and requires that only a fixed set of events can have zero probabilities (such as logical contradictions), but then the geometry of reason ends up in the catch-22 of LP conditioning running afoul of REGULARITY.

LP conditioning violates EXPANSIBILITY.

4.6 Horizon

One example for the horizon effect is George Schlesinger's comparison between the risk of a commercial airplane crash and the risk of a military glider landing in enemy territory.

> **Example 6: Airplane Gliders.** Compare two scenarios. In the first, an airplane which is considered safe (probability of crashing is $1/10^9$) goes through an inspection where a mechanical problem is found which increases the probability of a crash to $1/100$. In the second, military gliders land behind enemy lines, where their risk of perishing is 26%. A slight change in weather pattern increases this risk to 27%. (Schlesinger, 1995, 211)

I claim that an amujus ought to fulfill the requirements of the horizon effect: it ought to be more difficult to update as probabilities become more extreme (or less middling). I have formalized this requirement in **List B**. It is trivial that the geometry of reason does not fulfill it. Information theory fails as well, which gives the horizon effect its prominent place in both lists. The way information theory fails, however, is quite different. Near the boundary of $\mathbb{S}^{n-1}$, information theory reflects the horizon effect just as our expectation requires. The problem is near the centre, where some equidistant points are more divergent the closer they are to the middle. I will give an example and more explanation in subsection 5.2.

5 Expectations for Information Theory

Asymmetry is the central feature of the difference concept that information theory proposes for the purpose of updating between finite probability distributions. In information theory, the information loss differs depending on whether one uses probability distribution P to encode a message distributed according to probability distribution Q, or whether one uses probability distribution Q to encode a message distributed according to probability distribution P. This asymmetry may very well carry over into the epistemic realm. Updating from one probability distribution, for example, which has $P(X) = x > 0$ to $P'(X) = 0$ is common. It is called standard conditioning. Going in the opposite direction, however, from $P(X) = 0$ to $P'(X) = x' > 0$ is controversial and unusual.

The Kullback-Leibler divergence, which is the most promising concept of difference for probability distributions in information theory and the one which gives us Bayesian standard conditioning as well as Jeffrey conditioning, is non-commutative and may provide the kind of asymmetry required to reflect epistemic asymmetry. However, it also violates TRIANGULARITY, COLLINEAR HORIZON, and TRANSITIVITY OF ASYMMETRY. The task of this section is to show how serious these violations are.

5.1 Triangularity

The three points A, B, C in (2) violate TRIANGULARITY:

$$D_{\mathrm{KL}}(A, C) > D_{\mathrm{KL}}(B, C) + D_{\mathrm{KL}}(A, B). \tag{25}$$

This is counterintuitive on a number of levels, some of which I have already hinted at in illustration: taking a shortcut while making a detour; buying a pair of shoes for more money than buying the shoes individually.

Information theory, however, does not only violate TRIANGULARITY. It violates it in a particularly egregious way. Consider any distinct two points x

and z on $\mathbb{S}^{n-1}$ with coordinates x_i and z_i ($1 \leq i \leq n$). For simplicity, let us write $\delta(x,z) = D_{\text{KL}}(z,x)$. Then, for any $\vartheta \in (0,1)$ and an intermediate point y with coordinates $y_i = \vartheta x_i + (1-\vartheta)z_i$, the following inequality holds true:

$$\delta(x,z) > \delta\left(x,y\right) + \delta\left(y,z\right). \tag{26}$$

I will prove this in a moment, but here is a disturbing consequence: think about an ever more finely grained sequence of partitions y^j, $j \in \mathbb{N}$, of the line segment from x to z with y^{jk} as dividing points. I will spare myself defining these partitions, but note that any dividing point y^{j_0k} will also be a dividing point in the more finely grained partitions y^{jk} with $j \geq j_0$. Then define the sequence

$$T_j = \sum_k \delta\left(y^{jk}, y^{j(k+1)}\right) \tag{27}$$

such that the sum has as many summands as there are dividing points for j, plus one (for example, two dividing points divide the line segment into three possibly unequal thirds). If δ were the Euclidean distance norm, T_j would be constant and would equal $\|z-x\|$. Zeno's arrow moves happily along from x to z, no matter how many stops it makes on the way. Not so for information theory and the Kullback-Leibler divergence. According to (26), any stop along the way reduces the sum of divergences.

T_j is a strictly decreasing sequence (does it go to zero? – I do not know, but if yes, it would add to the poignancy of this violation). The more stops you make along the way, the closer you bring together x and z.

For the proof of (26), it is straightforward to see that (26) is equivalent to

$$\sum_{i=1}^{n}(z_i - x_i)\log\frac{\vartheta x_i + (1-\vartheta)z_i}{x_i} > 0. \tag{28}$$

Now I use the following trick. Expand the right hand side to

$$\sum_{i=1}^{n}\left(z_i+\frac{\vartheta}{1-\vartheta}x_i-\frac{\vartheta}{1-\vartheta}x_i-x_i\right)\log\frac{\frac{1}{1-\vartheta}\left(\vartheta x_i+(1-\vartheta)z_i\right)}{\frac{1}{1-\vartheta}x_i}>0. \quad (29)$$

(29) is clearly equivalent to (28). It is also equivalent to

$$\sum_{i=1}^{n}\left(z_i+\frac{\vartheta}{1-\vartheta}x_i\right)\log\frac{z_i+\frac{\vartheta}{1-\vartheta}x_i}{\frac{1}{1-\vartheta}x_i}+\sum_{i=1}^{n}\frac{1}{1-\vartheta}x_i\log\frac{\frac{1}{1-\vartheta}x_i}{z_i+\frac{\vartheta}{1-\vartheta}x_i}>0, \quad (30)$$

which is true by Gibbs' inequality.

5.2 Collinear Horizon

There are two intuitions at work that need to be balanced: on the one hand, the geometry of reason is characterized by simplicity, and the lack of curvature near extreme probabilities may be a price worth paying; on the other hand, simple examples such as example 6 make a persuasive case for curvature.

Information theory is characterized by a very complicated 'semi-quasimetric' (the attribute 'quasi' is due to its non-commutativity, the attribute 'semi' to its violation of the triangle inequality). One of its initial appeals is that it performs well with respect to the horizon requirement near the boundary of the simplex, which is also the location of Schlesinger's examples. It is not trivial, however, to articulate what the horizon requirement really demands.

COLLINEAR HORIZON in **List B** seeks to set up the requirement as weakly as possible, only demanding that points collinear with the centre exhibit the horizon effect. The hope is that continuity will take care of the rest, since I want the horizon effect also for probability distributions that are not collinear with the centre. Be that as it may, the Kullback-Leibler divergence fails COLLINEAR HORIZON. Here is a simple example.

$$p=\left(\frac{1}{5},\frac{2}{5},\frac{2}{5}\right) \qquad p'=q=\left(\frac{1}{4},\frac{3}{8},\frac{3}{8}\right) \qquad q'=\left(\frac{3}{10},\frac{7}{20},\frac{7}{20}\right) \quad (31)$$

The conditions of COLLINEAR HORIZON in **List B** are fulfilled. If p represents A, p' and q represent B, and q' represents C, then note that $\|B - A\| = \|C - B\|$ and M, A, B, C are collinear. In violation of COLLINEAR HORIZON,

$$D_{\mathrm{KL}}(B, A) = 7.3820 \cdot 10^{-3} > 6.4015 \cdot 10^{-3} = D_{\mathrm{KL}}(C, B). \tag{32}$$

This violation of an expectation is not as serious as the violation of TRIANGULARITY or TRANSITIVITY OF ASYMMETRY. It is opaque, however, what motivates information theory not only to put probability distributions farther apart near the periphery, as I would expect, but also near the centre. I lack the epistemic intuition reflected in the behaviour (for graphical illustration see figure 5 and figure 6). The next subsection on asymmetry deals with this lack of epistemic intuition writ large.

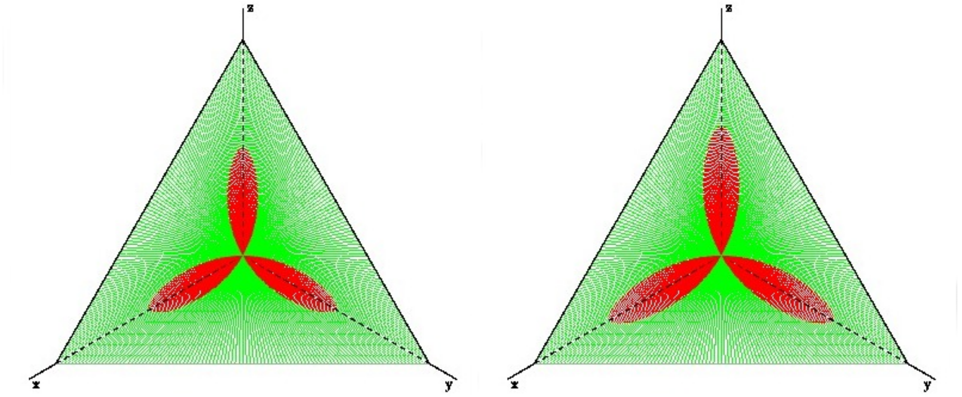

Figure 5: The diagram on the left displays all points in red which violate COLLINEAR HORIZON for information theory, measured from the centre. The diagram on the right displays points in different colours whose orientation of asymmetry differs, also for information theory, measured from the centre. The two red sets are not the same, but there appears to be a relationship, one that ultimately I suspect to be due to the more basic property of asymmetry. What motivates the particular shape of these sets is not clear. There is not a simple intuition at work as there is for the geometry of reason.

5.3 Transitivity of Asymmetry

Recall Joyce's two axioms Weak Convexity and Symmetry (see page 4). The geometry of reason (certainly in its Euclidean form) mandates Weak

Convexity because the bisector of an isosceles triangle is always shorter than the isosceles sides. Weak Convexity also holds for information theory. Symmetry, however, fails for information theory. Fortunately, although I do not pursue this any further here, information theory arrives at many of Joyce's results even without the violated axiom.

Asymmetry presents a problem for the geometry of reason as well as for information theory. For the geometry of reason, the problem is akin to CONTINUITY. For information theory, the problem is the non-trivial nature of the asymmetries it induces, which somehow need to be reconnected to epistemic justification. I will consider this problem in a moment, but first I will have a look at the problem for the geometry of reason.

Extreme probabilities are special and create asymmetries in updating: moving in direction from certainty to uncertainty is asymmetrical to moving in direction from uncertainty to certainty. Geometry of reason's metric topology, however, allows for no asymmetries.

> **Example 7: Extreme Asymmetry.** Consider two cases where for case 1 the prior probabilities are $Y_1 = (0.4, 0.3, 0.3)$ and the posterior probabilities are $Y_1' = (0, 0.5, 0.5)$; for case 2 the prior probabilities are reversed, so $Y_2 = (0, 0.5, 0.5)$ and the posterior probabilities $Y_2' = (0.4, 0.3, 0.3)$.

Case 1 is a straightforward application of standard conditioning. Case 2 is more complicated: what does it take to raise a prior probability of zero to a positive number? In terms of information theory, the information required is infinite. Case 2 is also not compatible with standard conditioning (at least not with what Alan Hájek calls the ratio analysis of conditional probability, see Hájek, 2003). The geometry of reason may want to solve this problem by signing on to a version of regularity, but then it violates REGULARITY. Happy kids, clean house, sanity: the hapless homemaker must pick two. The third remains elusive. Continuity, a consistent view of regularity, and symmetry: the hapless geometer of reason cannot have it all.

Now turn to the woes of the information theorist. Given the asymmetric similarity measure of probability distributions that information theory requires (the Kullback-Leibler divergence), a prior probability distribution P may be closer to a posterior probability distribution Q than Q is to P if their roles (prior-posterior) are reversed. That is just what we would expect. The problem is that there is another posterior probability distribution R where the situation is just the opposite: prior P is further away from posterior R

than prior R is from posterior P. And whether a probability distribution different from P is of the Q-type or of the R-type escapes any epistemic intuition.

For simplicity, let us consider probability distributions and their associated credence functions on an event space with three atoms $\Omega = \{\omega_1, \omega_2, \omega_3\}$. The simplex $\mathbb{S}^2$ represents all of these probability distributions. Every point p in $\mathbb{S}^2$ representing a probability distribution P induces a partition on $\mathbb{S}^2$ into points that are symmetric to p, positively skew-symmetric to p, and negatively skew-symmetric to p given the topology of information theory.

In other words, if

$$\Delta_P(P') = D_{\text{KL}}(P', P) - D_{\text{KL}}(P, P'), \tag{33}$$

then, holding P fixed, $\mathbb{S}^2$ is partitioned into three regions,

$$\Delta^{-1}(\mathbb{R}_{>0}) \qquad \Delta^{-1}(\mathbb{R}_{<0}) \qquad \Delta^{-1}(\{0\}) \tag{34}$$

One could have a simple epistemic intuition such as ‘it takes less to update from a more uncertain probability distribution to a more certain probability distribution than the reverse direction,’ where the degree of certainty in a probability distribution is measured by its entropy. This simple intuition accords with what we said about extreme probabilities and it holds true for the asymmetric distance measure defined by the Kullback-Leibler divergence in the two-dimensional case where Ω has only two elements.

In higher-dimensional cases, however, the tripartite partition (34) is non-trivial—some probability distributions are of the Q-type, some are of the R-type, and it is difficult to think of an epistemic distinction between them that does not already presuppose information theory (see figure 6 for illustration).

On any account of well-behaved and ill-behaved asymmetries, the Kullback-Leibler divergence is ill-behaved. Of the four axioms as listed by Ralph Kopperman for a distance measure d (see Kopperman, 1988, 89), the Kullback-Leibler divergence violates both symmetry and triangularity, making it a ‘semi-quasimetric’:

(m1) $d(x, x) = 0$

(m2) $d(x, z) \leq d(x, y) + d(y, z)$ (triangularity)

(m3) $d(x, y) = d(y, x)$ (symmetry)

(m4) $d(x, y) = 0$ implies $x = y$ (separation)

The Kullback-Leibler divergence not only violates symmetry and triangularity, but also TRANSITIVITY OF ASYMMETRY. For a description of TRANSITIVITY OF ASYMMETRY see **List B**. For an example of it, consider

$$P_1 = \left(\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\right) \qquad P_2 = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right) \qquad P_3 = \left(\frac{2}{5}, \frac{2}{5}, \frac{1}{5}\right) \quad (35)$$

In the terminology of TRANSITIVITY OF ASYMMETRY in **List B**, (P_1, P_2) is asymmetrically positive, and so is (P_2, P_3). The reasonable expectation is that (P_1, P_3) is asymmetrically positive by transitivity, but for the example in (35) it is asymmetrically negative.

How counterintuitive this is (epistemically and otherwise) is demonstrated by the fact that in MDS (the multi-dimensional scaling of distance relationships) almost all asymmetric distance relationships under consideration are asymmetrically transitive in this sense, for examples see international trade in Chino, 1978; journal citation in Coombs, 1964; car switch in Harshman et al., 1982; telephone calls in Harshman and Lundy, 1984; interaction or input-output flow in migration, economic activity, and social mobility in Coxon, 1982; flight time between two cities in Gentleman et al., 2006, 191; mutual intelligibility between Swedish and Danish in van Ommen et al., 2013, 193; Tobler's wind model in Tobler, 1975; and the cyclist lovingly hand-sketched in Kopperman, 1988, 91.

This 'ill behaviour' of information theory begs for explanation, or at least classification (it would help, for example, to know that all reasonable non-commutative difference measures used for updating are ill-behaved). Kopperman's objective is primarily to rescue continuity, uniform continuity, Cauchy sequences, and limits for topologies induced by difference measures which violate triangularity, symmetry, and/or separation. Kopperman does not touch axiom (m1), while in the psychological literature (see especially Tversky, 1977) self-similarity is an important topic. This is why an initially promising approach to asymmetric modeling in Hilbert spaces by Chino (see Chino, 1978; Chino, 1990; Chino and Shiraiwa, 1993; and Saburi and

Chino, 2008) will not help us to distinguish well-behaved and ill-behaved asymmetries between probability distributions.

For a future research project, it would be lovely either to see information theory debunked in favour of an alternative geometry (this paper has demonstrated that this alternative will not be the geometry of reason); or to see uniqueness results for the Kullback-Leibler divergence to show that despite its ill behaviour the Kullback-Leibler is the right asymmetric distance measure on which to base inference and updating. Chentsov's theory of monotone invariance and Amari's theory of α-connections are potential candidates to provide such results as well as an epistemic justification for information theory.

6 Conclusion

Leitgeb and Pettigrew's reasoning to establish LP conditioning on the basis of the geometry of reason is valid. Given the failure of LP conditioning with respect to expectations in **List A**, it cannot be sound. The premise to reject is the geometry of reason. A competing approach, information theory, yields results that fulfill all of these expectations except HORIZON. Information theory, however, fails two other expectations identified in **List B**—expectations which the geometry of reason fulfills. I am left with loose ends and ample opportunity for further work. The epistemic utility approach, itself a relatively recent phenomenon, needs to come to a deeper understanding of its relationship with information theory. It is an open question, for example, if it is possible to provide a complete axiomatization consistent with information theory to justify probabilism, standard conditioning, and Jeffrey conditioning from an epistemic utility approach as Shore and Johnston have done from a pragmatic utility approach. It is also an open question, given the results of this paper, if there is hope reconciling information theory with intuitions we have about epistemic utility and its attendant quantitative concept of difference for partial beliefs.

References

Amari, Shun-ichi. *Differential-Geometrical Methods in Statistics.* Berlin, Germany: Springer, 1985.

Armendt, Brad. "Is There a Dutch Book Argument for Probability Kinematics?" *Philosophy of Science* 47, 4: (1980) 583–588.

Chino, Naohito. "A Graphical Technique for Representing the Asymmetric Relationships Between N Objects." *Behaviormetrika* 5, 5: (1978) 23–44.

———. "A Generalized Inner Product Model for the Analysis of Asymmetry." *Behaviormetrika* 17, 27: (1990) 25–46.

Chino, Naohito, and Kenichi Shiraiwa. "Geometrical Structures of Some Non-Distance Models for Asymmetric MDS." *Behaviormetrika* 20, 1: (1993) 35–47.

Coombs, Clyde H. *A Theory of Data.* New York, NY: Wiley, 1964.

Coxon, Anthony. *The User's Guide to Multidimensional Scaling.* Exeter, NH: Heinemann Educational Books, 1982.

Csiszár, Imre, and Paul C Shields. *Information Theory and Statistics: A Tutorial.* Hanover, MA: Now Publishers, 2004.

Diaconis, Persi, and Sandy Zabell. "Updating Subjective Probability." *Journal of the American Statistical Association* 77, 380: (1982) 822–830.

Gentleman, R., B. Ding, S. Dudoit, and J. Ibrahim. "Distance Measures in DNA Microarray Data Analysis." In *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, edited by R. Gentleman, V. Carey, W. Huber, R. Irizarry, and S. Dudoit, Springer, 2006.

Hájek, Alan. "What Conditional Probability Could Not Be." *Synthese* 137, 3: (2003) 273–323.

Harshman, Richard, and Margaret Lundy. "The PARAFAC Model for Three-Way Factor Analysis and Multidimensional Scaling." In *Research methods for multimode data analysis*, edited by Henry G. Law, New York, NY: Praeger, 1984, 122–215.

Harshman, Richard A., Paul E. Green, Yoram Wind, and Margaret E. Lundy. "A Model for the Analysis of Asymmetric Data in Marketing Research." *Marketing Science* 1, 2: (1982) 205–242.

Howson, Colin, and Allan Franklin. "Bayesian Conditionalization and Probability Kinematics." *The British Journal for the Philosophy of Science* 45, 2: (1994) 451–466.

Jeffrey, Richard. *The Logic of Decision.* New York, NY: McGraw-Hill, 1965.

Joyce, James. "A Nonpragmatic Vindication of Probabilism." *Philosophy of Science* 65, 4: (1998) 575–603.

Kopperman, Ralph. "All Topologies Come from Generalized Metrics." *American Mathematical Monthly* 95, 2: (1988) 89–97.

Leitgeb, Hannes, and Richard Pettigrew. "An Objective Justification of Bayesianism II: The Consequences of Minimizing Inaccuracy." *Philosophy of Science* 77, 2: (2010) 236–272.

Levinstein, Benjamin Anders. "Leitgeb and Pettigrew on Accuracy and Updating." *Philosophy of Science* 79, 3: (2012) 413–424.

Lukits, Stefan. "Maximum Entropy and Probability Kinematics Constrained by Conditionals." *Entropy* 17, Special Issue "Maximum Entropy Applied to Inductive Logic and Reasoning," edited by Jürgen Landes and Jon Williamson: (2015) 1690–1700.

MacKay, David. *Information Theory, Inference and Learning Algorithms.* Cambridge, UK: Cambridge, 2003.

Miller, David. "A Geometry of Logic." In *Aspects of Vagueness*, edited by Heinz Skala, Settimo Termini, and Enric Trillas, Dordrecht, Holland: Reidel, 1984, 91–104.

Mormann, Thomas. "Geometry of Logic and Truth Approximation." *Poznan Studies in the Philosophy of the Sciences and the Humanities* 83, 1: (2005) 431–454.

van Ommen, Sandrien, Petra Hendriks, Dicky Gilbers, Vincent van Heuven, and Charlotte Gooskens. "Is Diachronic Lenition a Factor in the Asymmetry in Intelligibility Between Danish and Swedish?" *Lingua* 137: (2013) 193–213.

Saburi, S., and N. Chino. "A Maximum Likelihood Method for an Asymmetric MDS Model." *Computational Statistics & Data Analysis* 52, 10: (2008) 4673–4684.

Schlesinger, George. "Measuring Degrees of Confirmation." *Analysis* 55, 3: (1995) 208–212.

Shore, J., and R.W. Johnson. "Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-Entropy." *IEEE Transactions on Information Theory* 26, 1: (1980) 26–37.

Skyrms, Brian. "Dynamic Coherence." In *Advances in the Statistical Sciences: Foundations of Statistical Inference*, Springer, 1986, 233–243.

Tobler, Waldo. *Spatial Interaction Patterns.* Schloss Laxenburg, Austria: International Institute for Applied Systems Analysis, 1975.

Tversky, Amos. "Features of Similarity." *Psychological Review* 84, 4: (1977) 327–352.

Wagner, Carl. "Probability Kinematics and Commutativity." *Philosophy of Science* 69, 2: (2002) 266–278.

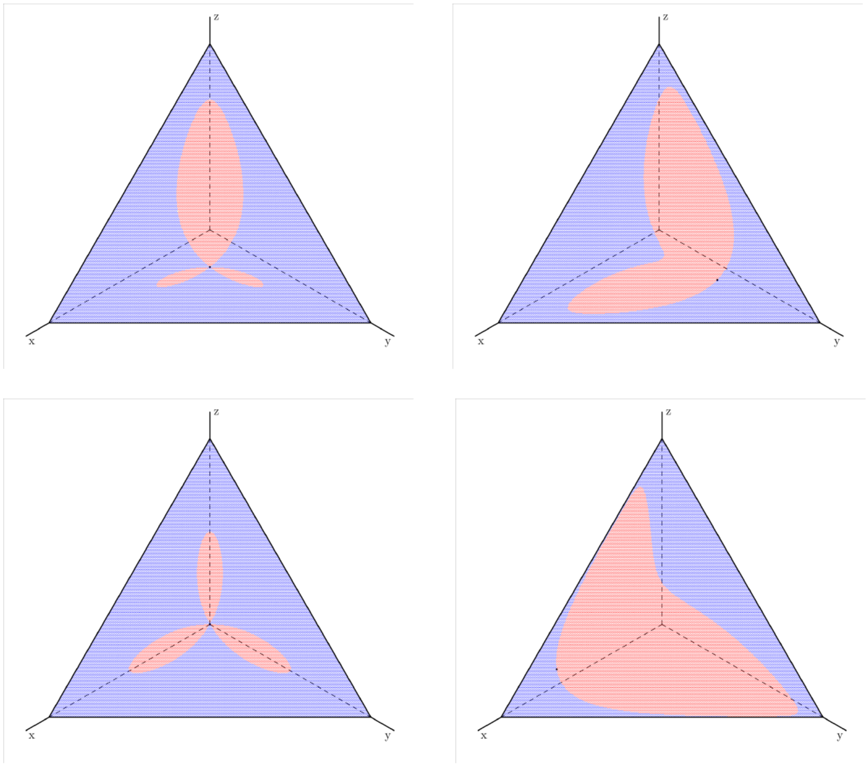

Figure 6: The partition (34) based on different values for P. From top left to bottom right, $P = (0.4, 0.4, 0.2); P = (0.242, 0.604, 0.154); P = (1/3, 1/3, 1/3); P = (0.741, 0.087, 0.172)$. Note that for the geometry of reason, the diagrams are trivial. The challenge for information theory is to explain the non-triviality of these diagrams epistemically without begging the question.