Asymmetry and the Geometry of Reason

Stefan Lukits

1 Introduction

Consider an idealized rational agent whose doxastic state at time t can be characterized such that she entertains certain partial beliefs with respect to a finite propositional algebra (the result of a die roll, for example). At a later time t' her beliefs may have changed due to intervening evidence. This paper presents a few results about how a rational agent's updated partial beliefs may be justified in terms of epistemic utility.

The agent uses ampliative reasoning to narrow down the set of logically and probabilistically coherent updates to a strict subset of credences (a so-called credal state). In Bayesian epistemology, for example, this credal state depends in some way on standard conditioning (formally expressed as $P(B|A)$) if the rational agent observes an event A between t and t'. There are numerous ways in which justification has been provided for this kind of ampliative reasoning. The probabilistically coherent partial beliefs may be a strict subset of logically coherent partial beliefs, so the first step of justification often consists in showing that partial beliefs are probabilities

and obey the axioms of probability theory.

In additional steps of justification, it is sometimes shown that based on certain assumptions the updated partial beliefs are in accordance with Bayes' theorem; or Jeffrey conditioning; or the principle of minimum cross-entropy. I will explain some of the formal apparatus behind these additional steps below. Before I go into such detail, however, I want to distinguish between types of justification. Often, these types work hand in hand with each other and are not mutually exclusive. Frank Ramsey and Bruno de Finetti are the pioneers of an approach that emphasizes pragmatic justification (see Ramsey, 1926; and de Finetti, 1931). A rational agent's partial beliefs should not be such that they can get the rational agent entangled in an arbitrage scheme run against her. Bas van Fraassen and Abner Shimony give arguments that the updated credences of a rational agent are of a certain kind because the agent wants her partial beliefs to cohere with her estimation of relative frequencies; in other words, she wants to be well-calibrated (see van Fraassen, 1983; and Shimony, 1988).

Harold Jeffreys and E.T. Jaynes try something different, based on an idea by John Maynard Keynes (see Keynes, 1921; Jeffreys, 1939; Jaynes and Bretthorst, 1998): a logic of belief change that is modeled on the logic of full beliefs. In this logic, prior beliefs plus intervening evidence necessitate the posterior beliefs much like the conclusion of an argument is necessitated by the argument's premises. The assumptions necessary to enable this flow of updating and, in certain cases, to provide unique results, strike many as overbearing and may even produce counterintuitive partial beliefs.

James Joyce, in his paper "A Nonpragmatic Vindication of Probabilism," uses an approach based on epistemic utility rather than pragmatic utility or logic of partial belief change, see Joyce, 1998). For Joyce, norms of gradational accuracy characterize the epistemic utility approach to partial beliefs, analogous to norms of truth for full beliefs. For this type of justification (let us call it more narrowly 'epistemic justification'), the rational agent's ampliative reasoning renders inadmissible all of those partial beliefs that are demonstrably inferior in epistemic terms to those partial beliefs left in the updated credal state.

A norm of truth for full beliefs would compare sets such as believed truths, believed falsehoods, disbelieved truths, disbelieved falsehoods, as well as truths and falsehoods that are neither believed nor disbelieved and derive some measure of epistemic utility from them. Often, this measure will not give a determinate answer to the question which beliefs are more epistemically virtuous—sometimes, these kinds of questions can be answered only by considering pragmatics, for example when a small number of false negatives is more damaging than a large number of false negatives (see Stephens, 2001). At other times, however, a set of full beliefs can be genuinely defective in epistemic terms compared to another set of full beliefs, for example when more falsehoods are believed and all else is equal.

By analogy, a set of partial beliefs can be genuinely defective in epistemic terms compared to another set of partial beliefs, for example when all of its truth estimates are worse than all of the truth estimates of its rival. Note that with an estimate, as opposed to a guess, I am not simply right or wrong

about the estimated quantity—I am instead more or less accurate. This distinction is due to Richard Jeffrey (see Jeffrey, 1986). Joyce uses his norm of gradational accuracy together with six axioms (structure, extensionality, dominance, normality, weak convexity, symmetry) to give an epistemic justification of probabilism: the requirement for a rational agent to keep her partial beliefs in keeping with the axioms of probability theory.

It is a natural question to ask whether the same line of reasoning can give us an epistemic justification of standard conditioning and Jeffrey conditioning. The former updates partial beliefs in light of an event that is known to be true based on intervening evidence, i.e. $P'(E) = 1$ (I am using P' for the posterior already with the assumption that the credence is a probability; the use of the expression 'know to be true' is not meant to make light of more intricate accounts about the relationship between knowledge and credences). The latter updates partial beliefs in light of an event about which the agent has shifted in uncertainty based on intervening evidence, i.e. $P'(E) = y$ where y is not necessarily equal to $P(E) = x$, the prior of E. In a series of articles that will be pivotal for the rest of this paper, Hannes Leitgeb and Richard Pettigrew show that using Joyce's approach, accuracy can be bifurcated into local and global accuracy. For this bifurcation to give consistent results, the Brier score must be used for Joyce's norm of gradational accuracy. The Brier score vindicates standard conditioning and rules out Jeffrey conditioning. A new type of conditioning, which I shall call LP conditioning, takes the place of Jeffrey conditioning (for details see Leitgeb and Pettigrew, 2010a).

I will show that LP conditioning fails a host of expectations that are reasonable to have for the kind of updating scenario that LP conditioning addresses. Since Leitgeb and Pettigrew's reasoning is valid, it cannot be sound. I identify a premise and call it the geometry of reason, on which Leitgeb and Pettigrew unwittingly cast doubt by reductio. The geometry of reason assumes that probability distributions entertain a geometric relationship to each other. At first, this assumption is natural: probability distributions can be isomorphically identified with points in a multi-dimensional simplex, which is a well-behaved $(n-1)$-dimensional subset of $\mathbb{R}^n$ (n is here the cardinality of the set of atoms belonging to the propositional algebra—that the credences on this set determine the credences on all propositions presumes probabilism). It seems natural to measure the difference between probability distributions by imposing a Euclidean metric on this space. Joyce uses this geometry of reason, for example, by defining midpoints between credences $(0.5P+0.5Q)$ and requiring them to be epistemically symmetric with respect to their parents P and Q (Joyce does not assume that they are probabilities, but seeks to justify it).

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)

If the geometry of reason gives us bad results, we need to look for alternatives. Joyce now appears to favour Bregman divergences (personal communication) and an emphasis on scoring rules rather than metrics (see Predd et al., 2009; for a connection between proper scoring rules and Bregman divergences see Abernethy and Frongillo, 2012). Bregman divergences are a generalization which encompasses Euclidean geometry (the squared Euclidean distance is a Bregman divergence) and its alternatives (for example the Kullback-Leibler divergence that I will mention below). Joyce, who in various places displays a preference for supervaluationist semantics, will say that one credence is epistemically inferior to another if it is inferior on all Bregman divergences. If different divergences give different results, inferior-

ity cannot be established.

I want to consider an alternative account based on Jaynes' approach. There is another way to relate probability distributions to each other and formulate a concept of difference between them. It is based on information theory, originally conceived to model the extra coding required when the sender gets the probability distribution of the alphabet not quite right. This special issue of *Synthese* wants to think about epistemic justification especially with respect to building a relationship between informal and formal epistemology. These elements come together nicely here: there is a substantial formal theory of information, serendipitously cast in terms of probabilities; there is a similarly substantial formal theory of probabilities and Bayesian epistemology, also cast in terms of probabilities. The formal theories are different: one uses Shannon's entropy, a logarithmic measure of difference; the other uses primarily ratios. It turns out that the superficial differences are rooted in deep commonalities. Information theory can serve as justification for Bayesian norms (probabilism, standard conditioning, and Jeffrey conditioning, for the latter two see Lukits, 2015, 1697f).

My claim is that this confluence of two formal theories has in itself justificatory force. This justificatory force may be small, and there are certainly detractors who claim that information theory sometimes, if not often, gives the wrong results if applied to epistemology. I myself in this paper will add to the wrinkles that appear when information theory and epistemology are wedded to each other. I will hopefully also show, however, that a Euclidean prejudice in these matters is incomparably more deleterious. Joyce's supervaluationist

solution to these problems using Bregman divergences is certainly a reasonable option. Just as there is something special about Euclidean geometry and its attendant Brier scoring rule, however, there is something special about information theory and its attendant Kullback-Leibler divergence. My results in this paper are largely negative: there are insurmountable problems for the geometry of reason; there are serious problems for information theory. The positive aspect is that formal accounts, when they sometimes show surprising coherence or when they can be used to yield informative theorems, can supply epistemic justification just in terms of their formal features.

1.1 Expectations for Jeffrey-Type Updating Scenarios

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 Joyce, 1998; Leitgeb and Pettigrew, 2010b; and Greaves and Wallace, 2006).

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). 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).

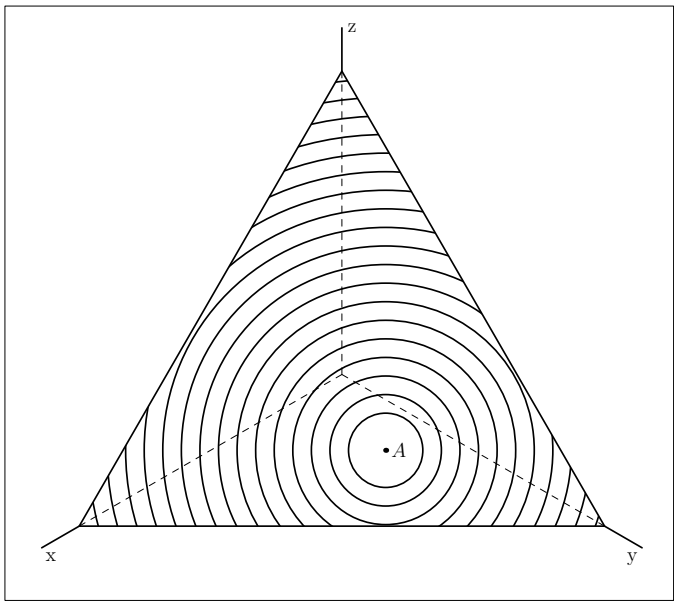

Figure 1: The simplex $\mathbb{S}^2$ in three-dimensional space $\mathbb{R}^3$ with contour lines corresponding to the geometry of reason around point A in equation (1). Points on the same contour line are equidistant from A with respect to the Euclidean metric. Compare the contour lines here to figure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

Consider the following example of a Jeffrey-type updating scenario.

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

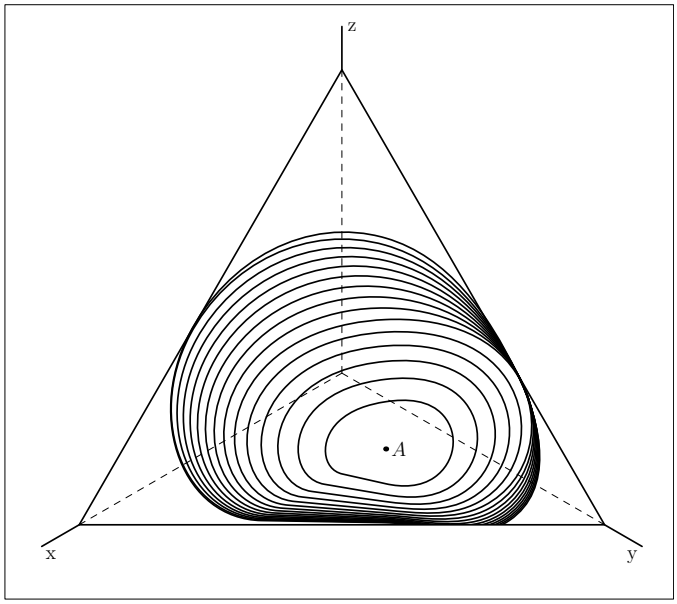

Figure 2: The simplex $\mathbb{S}^2$ with contour lines corresponding to information theory around point A in equation (1). Points on the same contour line are equidistant from A with respect to the Kullback-Leibler divergence. The contrast to figure 1 will become clear in much more detail in the body 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.

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. 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 quan-

titative 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.

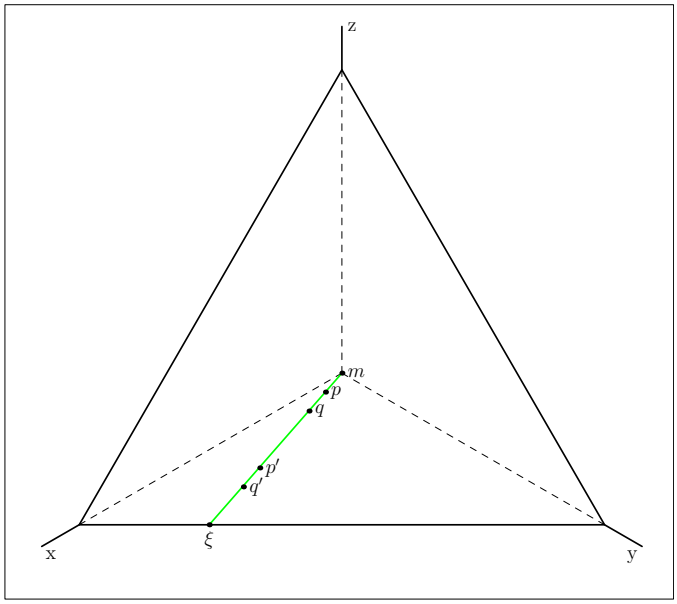

Figure 3: An illustrations of conditions (i)–(iii) for COLLINEAR HORIZON in **List B**. p, p' and q, q' must be equidistant and collinear with m and ξ. If q, q' is more peripheral than p, p', then COLLINEAR HORIZON requires that $|d(p, p')| < |d(q, q')|$.

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**. Conversely, 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 alsp 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 infor-

mation 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.

2 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:

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

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 mutually exclusive events. Now call $D_{\mathrm{KL}}(B, A)$ 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):

$$D_{\mathrm{KL}}(B, A) = \sum_{i=1}^{3} b_i \log \frac{b_i}{a_i}. \tag{2}$$

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:

$$\|B - A\| = \sqrt{\sum_{i=1}^{n} (b_i - a_i)^2}. \tag{3}$$

The Euclidean distance $\|B - A\|$ is defined as in equation (3). What is remarkable about the three points in (1) is that

$$\|C - A\| \approx 0.204 < \|B - A\| \approx 0.212 \tag{4}$$

and

$$D_{\mathrm{KL}}(B, A) \approx 0.0589 < D_{\mathrm{KL}}(C, A) \approx 0.069. \tag{5}$$

The Kullback-Leibler divergence and Euclidean distance give different recommendations with respect to proximity. In terms of the global inaccuracy measure presented in Leitgeb and Pettigrew (see Leitgeb and Pettigrew, 2010a, 206) and $E = W$ (all possible worlds are epistemically accessible),

$$\mathrm{GExp}_A(C) \approx 0.653 < \mathrm{GExp}_A(B) \approx 0.656. \tag{6}$$

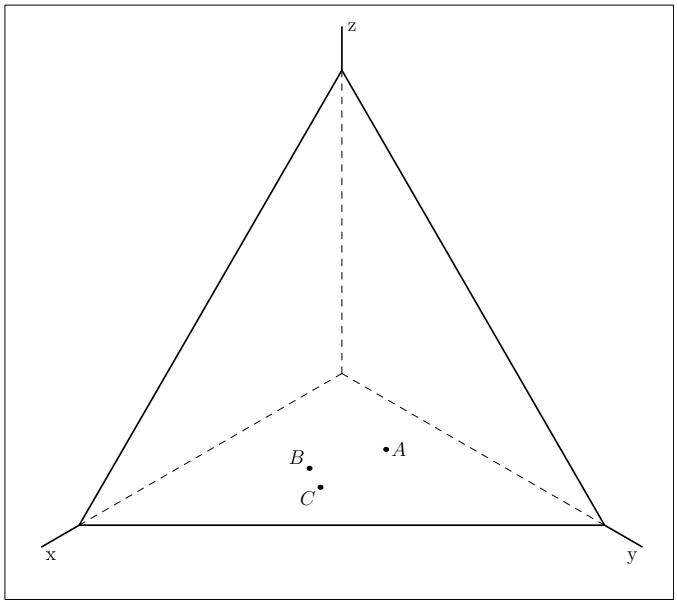

Figure 4: The simplex $\mathbb{S}^2$ in three-dimensional space $\mathbb{R}^3$ with points a, b, c as in equation (1) representing probability distributions A, B, C. Note that geometrically speaking C is closer to A than B is. Using the Kullback-Leibler divergence, however, B is closer to A than C is.

Global inaccuracy reflects the Euclidean proximity relation, not the recommendation of information theory. 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.

2.1 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

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

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,

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

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:

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

and then set

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

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

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

Compare these results to the results of Jeffrey conditioning:

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

Note that (7), (12), and (11) correspond to A, B, C in (1).

3 Expectations for the Geometry of Reason

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

3.1 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

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

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

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

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

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

LP conditioning violates CONTINUITY.

3.2 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:

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

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

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.

3.3 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.

3.4 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

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

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,

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

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

they do not agree if they use LP conditioning,

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

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

LP conditioning violates INVARIANCE.

3.5 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

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

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:

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

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.

3.6 Horizon

> **Example 6: Undergraduate Complaint.** An undergraduate student complains to the department head that the professor will not reconsider an 89% grade (which misses an A+ by one percent) when reconsideration was given to other students with a 67% grade (which misses a B- by one percent).

Intuitions may diverge, but the professor's reasoning is as follows. To improve a 60% paper by ten percent is easily accomplished: having your roommate check your grammar, your spelling, and your line of argument will sometimes do the trick. It is incomparably more difficult to improve an 85% paper by ten percent: it may take doing a PhD to turn a student who writes the former into a student who writes the latter. A maiore ad minus, the step from 89% to 90% is greater than the step from 67% to 68%.

Another 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 7: 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 4.2.

4 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.

4.1 Triangularity

There is an interesting connection between LP conditioning and Jeffrey conditioning as updating methods. Let B be on the zero-sum line between A and C if and only if

$$d(A, C) = d(A, B) + d(B, C) \tag{25}$$

where d is the difference measure we are using, so $d(A, B) = \|B - A\|$ for the geometry of reason and $d(A, B) = D_{\mathrm{KL}}(B, A)$ for information geometry. For the geometry of reason (and Euclidean geometry), the zero-sum line between two probability distributions is just what we intuitively think of as a straight line: in Cartesian coordinates, B is on the zero-sum line strictly between A and C if and only if for some $\vartheta \in (0, 1)$, $b_i = \vartheta a_i + (1 - \vartheta)c_i$ and $i = 1, \ldots, n$.

What the zero-sum line looks like for information theory is illustrated in figure 5. The reason for the oddity is that the Kullback-Leibler divergence does not obey TRIANGULARITY, an issue that I will address in detail in subsection 4.1). Call B a zero-sum point between A and C if (25) holds true. For the geometry of reason, the zero-sum points are simply the points on the straight line between A and C. For information geometry, the zero-sum points are the boundary points of the set where you can take a shortcut by making a detour, i.e. all points for which $d(A, B) + d(B, C) < d(A, C)$.

Remarkably, if A represents a relatively prior probability distribution and C the posterior probability distribution recommended by LP conditioning, the posterior probability distribution recommended by Jeffrey conditioning is always a zero-sum point with respect to the Kullback-Leibler divergence:

$$D_{\mathrm{KL}}(C, A) = D_{\mathrm{KL}}(B, A) + D_{\mathrm{KL}}(C, B) \tag{26}$$

Informationally speaking, if you go from A to C, you can just as well go from

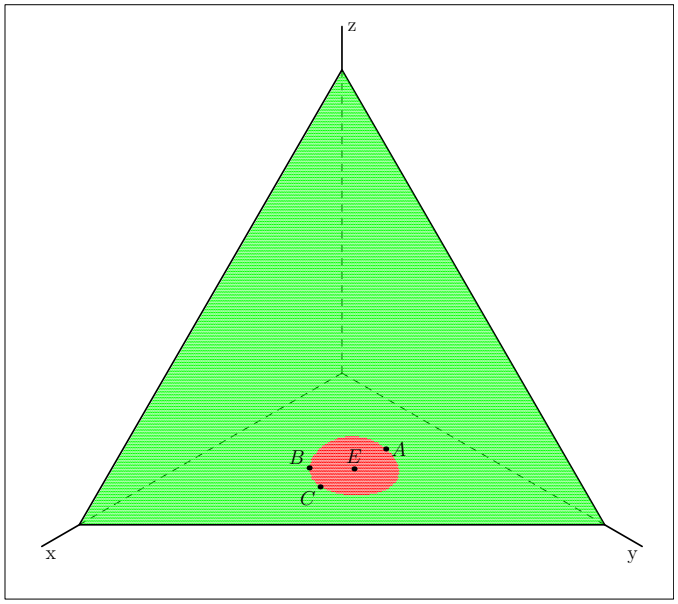

Figure 5: The zero-sum line between A and C is the boundary line between the green area, where the triangle inequality holds, and the red area, where the triangle inequality is violated. The posterior probability distribution B recommended by Jeffrey conditioning always lies on the zero-sum line between the prior A and the LP posterior C, as per equation (26). E is the point in the red area where the triangle inequality is most efficiently violated.

A to B and then from B to C. This does not mean that we can conceive of information geometry the way we would conceive of non-Euclidean geometry, where it is also possible to travel faster on what from a Euclidean perspective looks like a detour. For in information geometry, you can travel faster on what from the perspective of information theory (!) looks like a detour, i.e. the triangle inequality does not hold.

Before we get carried away with these analogies between divergences and metrics, however, it is important to note that it is not appropriate to im-

pose expectations that are conventional for metrics on divergences. Bregman divergences, for example, in some sense violate the triangle equality by design. If d_f is a Bregman divergence with the corresponding convex function f (for example, $f(x)$ is the inner product $\langle x, x\rangle$ for the squared Euclidean distance; or $f(x) = \sum_{i=1}^{n} x_i \log x_i$ for the Kullback-Leibler divergence), then for convex $C \in \mathbb{R}^n$ and all $x \in C$ and $y \in \mathbb{R}^n$ the following triangle inequality is true: $d_f(x, y) \geq d_f(x, y') + d_f(y', y)$, where y' is the projection of y onto C such that $d_f(z, y), z \in C$ is minimal. The squared Euclidean distance is an interesting case in point for this property (for obtuse triangles, $c^2 > a^2 + b^2$). To subject the difference concept between probability distribution to a TRIANGULARITY requirement may be a temptation to resist and only reveal another instance of Euclidean prejudice.

To prove equation (26) in the case $n = 3$ (assuming that LP conditioning does not 'fall off the edge' as in case (b) in Leitgeb and Pettigrew, 2010b, 253) note that all three points (prior, point recommended by Jeffrey conditioning, point recommended by LP conditioning) can be expressed using three variables:

$$\begin{aligned} A &= (1-\alpha, \beta, \alpha-\beta) \\ B &= \left(1-\gamma, \tfrac{\gamma\beta}{\alpha}, \tfrac{\gamma(\alpha-\beta)}{\alpha}\right) \\ C &= \left(1-\gamma, \beta + \tfrac{1}{2}(\gamma-\alpha), \alpha-\beta+\tfrac{1}{2}(\gamma-\alpha)\right) \end{aligned} \tag{27}$$

The rest is basic algebra using the definition of the Kullback-Leibler divergence in (2). To prove the claim for arbitrary n one simply generalizes (27). It is a handy corollary of (26) that whenever (A, B) and (B, C) violate TRANSITIVITY OF ASYMMETRY then

$$D_{\mathrm{KL}}(A, C) > D_{\mathrm{KL}}(B, C) + D_{\mathrm{KL}}(A, B) \tag{28}$$

Consequently, the three points A, B, C in (1) violate TRIANGULARITY. 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 $\delta(x, z) = D_{\mathrm{KL}}(z, x)$. 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:

$$\delta(x, z) > \delta\left(x, y\right) + \delta\left(y, z\right). \tag{29}$$

It is straightforward to see that (29) is equivalent to

$$\sum_{i=1}^{n} (z_i - x_i) \log \frac{\vartheta x_i + (1 - \vartheta) z_i}{x_i} > 0. \tag{30}$$

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

$$\sum_{i=1}^{n}\left(z_i+\frac{\vartheta}{1-\vartheta}x_i-\frac{\vartheta}{1-\vartheta}x_i-x_i\right)\log\frac{\frac{1}{1-\vartheta}\left(\vartheta x_i+(1-\vartheta)z_i\right)}{\frac{1}{1-\vartheta}x_i}>0. \quad (31)$$

(31) is clearly equivalent to (30). It is also equivalent to

$$\sum_{i=1}^{n}\left(z_i+\frac{\vartheta}{1-\vartheta}x_i\right)\log\frac{z_i+\frac{\vartheta}{1-\vartheta}x_i}{\frac{1}{1-\vartheta}x_i}+\sum_{i=1}^{n}\frac{1}{1-\vartheta}x_i\log\frac{\frac{1}{1-\vartheta}x_i}{z_i+\frac{\vartheta}{1-\vartheta}x_i}>0, \quad (32)$$

which is true by Gibbs' inequality. Like Bregman divergences in general, the Kullback-Leibler divergence in particular violates TRIANGULARITY by design.

4.2 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 7 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.

$$p = \left(\frac{1}{5}, \frac{2}{5}, \frac{2}{5}\right) \qquad p' = q = \left(\frac{1}{4}, \frac{3}{8}, \frac{3}{8}\right) \qquad q' = \left(\frac{3}{10}, \frac{7}{20}, \frac{7}{20}\right) \tag{33}$$

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,

$$D_{\mathrm{KL}}(B, A) = 7.3820 \cdot 10^{-3} > 6.4015 \cdot 10^{-3} = D_{\mathrm{KL}}(C, B). \tag{34}$$

This violation of an expectation is not as serious as the violation of TRIANGULARITY or TRANSITIVITY OF ASYMMETRY. Just as there is still a reasonable disagreement about difference measures (which do not exhibit the horizon effect) and ratio measures (which do) in degree of confirmation theory, most of us will not have strong intuitions about the adequacy of information

theory based on its violation of COLLINEAR HORIZON. One way in which I can attenuate the independent appeal of this violation against information theory is by making it parasitic on the asymmetry of information theory.

Figure 6 illustrates what I mean. Consider the following two inequalities, where M is represented by the centre m of the simplex with $m_i = 1/n$ and Y is an arbitrary probability distribution with X as the midpoint between M and Y, so $x_i = 0.5(m_i + y_i)$.

$$\text{(i) } D_{\mathrm{KL}}(Y, M) > D_{\mathrm{KL}}(M, Y) \text{ and (ii) } D_{\mathrm{KL}}(X, M) > D_{\mathrm{KL}}(Y, X) \tag{35}$$

In terms of coordinates, the inequalities reduce to

$$\text{(i) } H(y) < \frac{1}{n}\sum(\log y_i) - \log\frac{1}{n^2} \text{ and} \tag{36}$$

$$\text{(ii) } H(y) > \log\frac{4}{n} - \sum\left[\left(\frac{3}{2}y_i + \frac{1}{2n}\right)\log\left(y_i + \frac{1}{n}\right)\right]. \tag{37}$$

(i) is simply the case described in the next subsection for asymmetry and illustrated on the bottom left of figure 7. (ii) tells us how far from the midpoint I can go with a scenario where $p = m, p' = q$ while violating COLLINEAR HORIZON. Clearly, as illustrated in figure 6, there is a relationship

between asymmetry and COLLINEAR HORIZON.

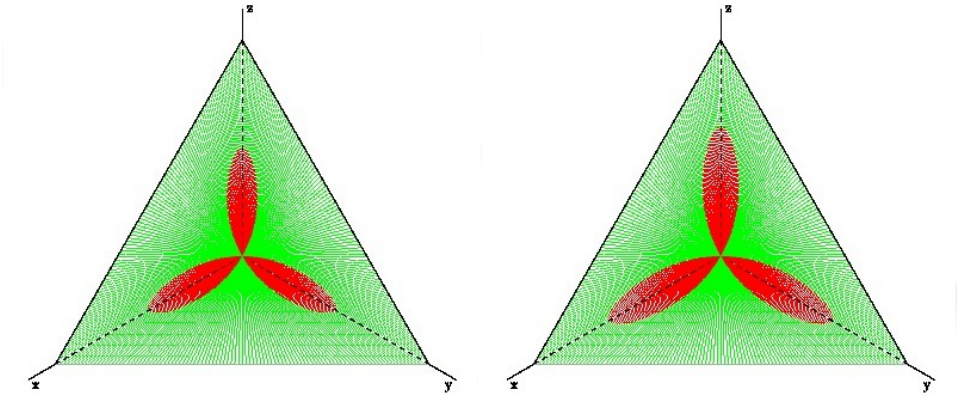

Figure 6: These two diagrams illustrate inequalities (36) and (37). The former displays all points in red which violate COLLINEAR HORIZON, measured from the centre. The latter displays points in different colours whose orientation of asymmetry differs, measured from the centre. The two red sets are not the same, but there appears to be a relationship, one that ultimately I suspect to be due to the more basic property of asymmetry.

It is opaque 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. The next subsection on asymmetry deals with this lack of epistemic intuition writ large.

4.3 Transitivity of Asymmetry

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 8: 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

$$\Delta_P(P') = D_{\mathrm{KL}}(P', P) - D_{\mathrm{KL}}(P, P'), \tag{38}$$

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

$$\Delta^{-1}(\mathbb{R}_{>0}) \qquad \Delta^{-1}(\mathbb{R}_{<0}) \qquad \Delta^{-1}(\{0\}) \tag{39}$$

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 (39) 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 7 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 triangu-

larity, but also TRANSITIVITY OF ASYMMETRY. For a description of TRANSITIVITY OF ASYMMETRY see **List B**. For an example of it, consider

$$P_1 = \left(\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\right) \qquad P_2 = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right) \qquad P_3 = \left(\frac{2}{5}, \frac{2}{5}, \frac{1}{5}\right) \quad (40)$$

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 (40) 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). For

a future research project, it would be interesting 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.