I am wondering who is your intended audience. This is quite a long and dense paper: for many conferences and journals it would be considered too long. The presentation of the ideas is suitable for specialist conferences or journals (where accessibility for a general philosophical audience is not a requirement), but not for general philosophy conferences or journals. This taken together with the length will limit your possible venues, but perhaps you already know (and are happy with) this!

Asymmetry and the Geometry of Reason

Stefan Lukits

Abstract

Defenders of the epistemic utility approach to Bayesian epistemology use the geometry of reason to justify the foundational Bayesian tenets of probabilism and standard conditioning. The geometry of reason is the view that the underlying topology for credence functions is a metric space, on the basis of which axioms and theorems of epistemic utility for partial beliefs are formulated. It implies that Jeffrey conditioning must cede to an alternative form of conditioning. The latter fails a long list of plausible expectations and implies unacceptable results in certain cases. One solution to this problem is to reject the geometry of reason and accept information theory in its stead. Information theory comes fully equipped with an axiomatic approach which covers probabilism, standard conditioning, and Jeffrey conditioning. It is not based on an underlying topology of a metric space, but uses non-commutative divergences instead of a symmetric distance measure. I show that information theory, despite great initial promise, also fails to accommodate basic epistemic intuitions.

1 Introduction

In the early 1970s, the dominant models for similarity in the psychological literature were all geometric in nature. Distance measures capturing similarity and dissimilarity between concepts obeyed minimality, symmetry, and the triangle inequality. Then Amos Tversky wrote a compelling paper undermining the idea that a metric topology is the best model (see Tversky, 1977). Tversky gave both theoretical and empirical reasons why similarity between concepts fulfills neither minimality, nor symmetry, nor the triangle inequality. Geometry with its metric distance measures was in some ways not a useful model of similarity. Tversky presented an alternative set-theoretic account which accommodated intuitions that could not be reconciled with a geometry of similarity.

The aim of this paper is to help along a similar paradigm shift when it comes to epistemic modeling of closeness or difference between subjective probability distributions. The 'geometry of reason' (a term coined by Richard Pettigrew and Hannes Leitgeb, two of its advocates) violates reasonable expectations for an acceptable model. A non-metric alternative, information theory, fulfills many of these expectations but violates others which are similarly intuitive. Instead of presenting a third alternative which coheres better with the list of expectations outlined in section 4, I defend the view that while the violations of the geometry of reason are irremediable, there is a promise in the wings that an advanced formal account of information theory, using the theory of differential manifolds, can explain information theory's violations of prima facie reasonable expectations.

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 $\mathbb{S}^5$, the five-dimensional simplex for which

$$p_1 + p_2 + p_3 + p_4 + p_5 + p_6 = 1. \tag{1}$$

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).

Epistemic utility in Bayesian epistemology has attracted some attention in the past few years. Patrick Maher provides a compelling acceptance-based account of epistemic utility (see Maher, 1993, 182–207). Joyce, in "A

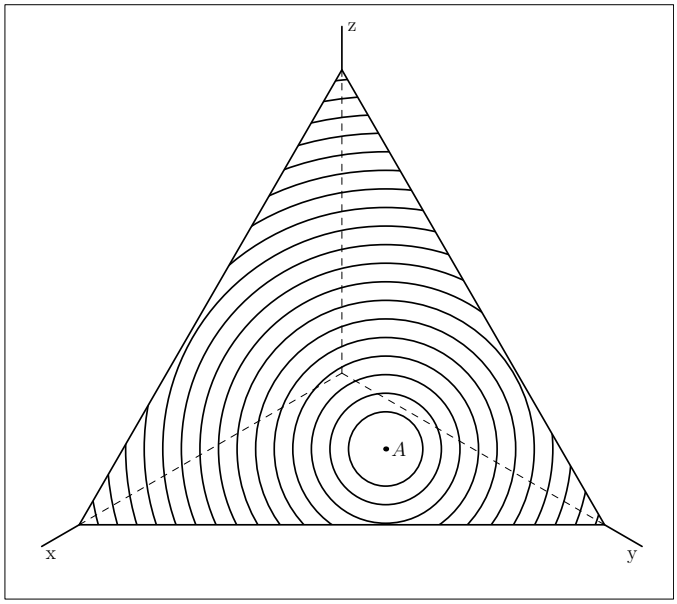

Figure 1: The simplex $\mathbb{S}^2$ in three-dimensional space $\mathbb{R}^3$ with contour lines corresponding to the geometry of reason around point A in equation (12). Points on the same contour line are equidistant from A with respect to the Euclidean metric. Compare the contour lines here to figure 2. Note that this diagram and all the following diagrams are frontal views of the simplex.

Nonpragmatic Vindication of Probabilism," defends probabilism supported by partial-belief-based epistemic utility rather than the pragmatic utility common in Dutch-book style arguments (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.

Wallace and Greaves investigate epistemic utility functions along 'stability' lines and conclude that for everywhere stable utility functions standard conditioning is optimal, while only somewhere stable utility functions create problems for maximizing expected epistemic utility norms (see Greaves and Wallace, 2006; and Pettigrew, 2013). Richard Pettigrew and Hannes Leitgeb have published arguments that under certain assumptions probabilism and standard conditioning (which together give epistemology a distinct Bayesian flavour) minimize inaccuracy, thereby providing maximal epistemic utility (see Leitgeb and Pettigrew, 2010a and 2010b).

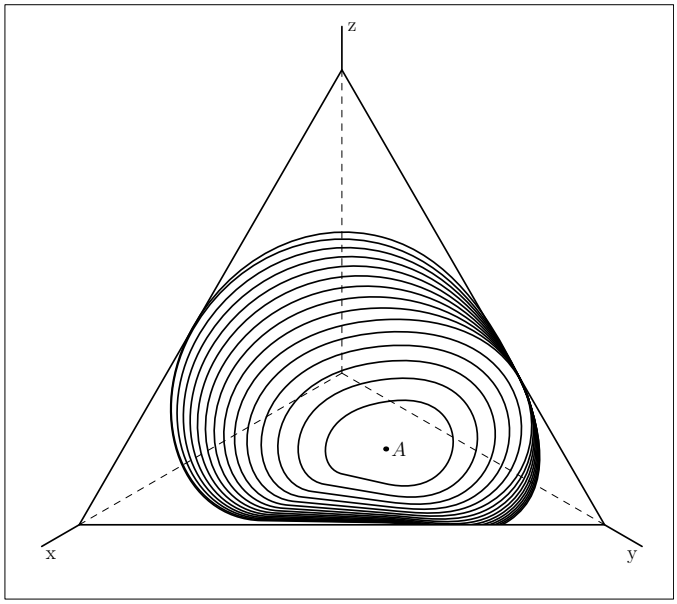

Figure 2: The simplex $\mathbb{S}^2$ with contour lines corresponding to information theory around point A in equation (12). Points on the same contour line are equidistant from A with respect to the Kullback-Leibler divergence. The contrast to figure 1 will become clear in much more detail in the body 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.

Leitgeb and Pettigrew show, given the geometry of reason and other axioms inspired by Joyce (for example normality and dominance), 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

$$\frac{1}{N}\sum_{i=1}^{N}(f_y - o_t) = \frac{1}{100}\left(32\cdot(0.3-1) + 68\cdot(0.3-0)\right) = 0.218 \qquad (2)$$

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 Pettigrew's account, using the Brier score, it fails to provide maximal epistemic utility. Another type of conditioning, which we 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, by reductio, doubt on the geometry of reason.

Unless co-authored, use "I"

Awkward phrasing

I will show that LP conditioning, which the geometry of reason entails, fails commonsense expectations that are reasonable to have for the kind of updating scenario that LP conditioning addresses. To relate probability distributions to each other geometrically, using the isomorphism between the set of probability distributions on a finite event space W with $|W| = n$ and the $n-1$-dimensional simplex $\mathbb{S}^{n-1} \subset \mathbb{R}^n$, is initially an arbitrary move. Leitgeb and Pettigrew do little to substantiate a link between the geometry of reason and epistemic utility on a conceptual level. It is the formal success of the model that makes the geometry of reason attractive, but the failure of LP conditioning to meet basic expectations undermines this success.

The question then remains whether we have a plausible candidate to supplant the geometry of reason. The answer is yes: information theory provides us with a measure of closeness between probability distributions on a finite event space that has more conceptual appeal than the geometry of reason, especially with respect to epistemic utility—it is intuitively correct to relate coming-to-knowledge to exchange of information. More persuasive than intuition, however, is the fact that information theory supports both standard conditioning (see Williams, 1980) and the extension of standard conditioning to Jeffrey conditioning (see Caticha and Giffin, 2006; and Lukits, 2015), an extension which is on the one hand intuitive (see Wagner, 2002) and on the other hand formally continuous with the standard conditioning which Leitgeb and Pettigrew have worked so hard to vindicate nonpragmatically. LP conditioning is not continuous with standard conditioning, which is reflected in one of the expectations that LP conditioning fails to meet.

2 Epistemic Utility and the Geometry of Reason

I would stick to talk of epistemic utility here rather than virtue, as "virtue epistemology" is its own enterprise (and you don't want to muddy the waters by getting into it here, I imagine).

2.1 Epistemic Utility for Partial Beliefs

There is more epistemic virtue for an agent in believing a truth rather than not believing it and in not believing a falsehood rather than believing it. Accuracy in full belief epistemology can be measured by counting four sets, believed truths and falsehoods as well as unbelieved truths and falsehoods, and somehow relating them to each other such that epistemic virtue is rewarded and epistemic vice penalized. Accuracy in partial belief epistemology must take a different shape since as a 'guess' all partial non-full beliefs are off the mark so that they need to be appreciated as 'estimates' instead. Richard Jeffrey distinguishes between guesses and estimates: a guess fails unless it is on target, whereas an estimate succeeds depending on how close it is to the target.

The gradational accuracy needed for partial belief epistemology is reminiscent of verisimilitude and its associated difficulties in the philosophy of science (see Popper, 1963; Gemes, 2007; and Oddie, 2013). Both Joyce and Leitgeb/Pettigrew propose 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 distributions 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 Euclidean structure. On the contrary, this would appear to be an improbable coincidence. (Mormann, 2005, 453)

2.2 Axioms for Epistemic Utility

Leitgeb and Pettigrew present the following salient axioms (see Leitgeb and Pettigrew, 2010a, 219):

> **Local Normality and Dominance**: If I is a legitimate inaccuracy measure, then there is a strictly increasing function $f : \mathbb{R}_0^+ \to \mathbb{R}_0^+$ such that, for any $A \in W$, $w \in W$, and $x \in \mathbb{R}_0^+$,
>
> $$I(A, w, x) = f\left(|\chi_A(w) - x|\right). \tag{3}$$

> **Global Normality and Dominance**: If G is a legitimate global inaccuracy measure, there is a strictly increasing function $g : \mathbb{R}_0^+ \to \mathbb{R}_0^+$ such that, for all worlds w and belief functions $b \in \mathrm{Bel}(W)$,
>
> $$G(w, b) = g\left(\|w - b_{\mathrm{glo}}\|\right). \tag{4}$$

Similarly to Joyce, these axioms are justified on the basis of geometry, but this time more explicitly so:

> Normality and Dominance [are] a consequence of taking seriously the talk of inaccuracy as 'distance' from the truth, and [they endorse] the geometrical picture provided by Euclidean n-space as the correct clarification of this notion. As explained in section 3.2, the assumption of this geometrical picture is one of the presuppositions of our account, and we do not have much to offer in its defense, except for stressing that we would be equally interested in studying the consequences of minimizing expected inaccuracy in a non-Euclidean framework. But without a doubt, starting with the Euclidean case is a natural thing to do.

Leitgeb and Pettigrew define two notions, local and global inaccuracy, and show that one must adopt a Brier score to measure inaccuracy in order to

avoid epistemic dilemmas trying to minimize inaccuracy on both measures. To give the reader an idea what this looks like in detail and for purposes of later exposition, I want to provide some of the formal apparatus. Let W be a set of worlds and $A \subseteq W$ a proposition. Then

$$I : P(W) \times W \times \mathbb{R}_0^+ \to \mathbb{R}_0^+ \tag{5}$$

is a measure of local inaccuracy such that $I(A, w, x)$ measures the inaccuracy of the degree of credence x with respect to A at world w. Let $\mathrm{Bel}(W)$ be the set of all belief functions (what we have been calling distributions of partial belief). Then

$$G : W \times \mathrm{Bel}(W) \to \mathbb{R}_0^+ \tag{6}$$

is a measure of global inaccuracy of a belief function b at a possible world w such that $G(w, b)$ measures the inaccuracy of a belief function b at world w.

Axioms such as normality and dominance guarantee that the only legitimate measure of inaccuracy are Brier scores if one wants to avoid epistemic dilemmas where one receives conflicting advice from the local and the global measures. For local inaccuracy measures, this means that there is $\lambda \in \mathbb{R}^+$ such that

$$I(A, w, x) = \lambda \left(\chi_A(w) - x\right)^2 \tag{7}$$

where χ_A is the characteristic function of A. For global inaccuracy measures, this means that there is $\mu \in \mathbb{R}^+$ such that

$$G(w, b) = \mu \|w - b\|^2 \tag{8}$$

where w and b are represented by vectors and $\|u - v\|$ is the Euclidean distance

$$\sqrt{\sum_{i=1}^{n} \left(u_i - v_i\right)^2}. \tag{9}$$

We use (7) to define expected local inaccuracy of degree of belief x in proposition A by the lights of belief function b, with respect to local inaccuracy measure I, and over the set E of epistemically possible worlds as follows:

$$\mathrm{LExp}_b(I, A, E, x) = \sum_{w \in E} b(\{w\}) I(A, w, x) = \sum_{w \in E} b(\{w\}) \lambda \left(\chi_A(w) - x\right)^2 . \tag{10}$$

We use (8) to define expected global inaccuracy of belief function b' by the lights of belief function b, with respect to global inaccuracy measure G, and over the set E of epistemically possible worlds as follows:

$$\mathrm{GExp}_b(G, E, b') = \sum_{w \in E} b(\{w\}) G(w, b') = \sum_{w \in E} b(\{w\}) \mu \|w - b\|^2. \tag{11}$$

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. Note that in terms of notation I for Joyce is global and related to Leitgeb and Pettigrew's G in (6) rather than I in (5).

> **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

Leitgeb and Pettigrew's work is continuous with Joyce's work, but significantly goes beyond it. Joyce wants much weaker assumptions and would be leery of expected inaccuracies (10) and (11), as they might presuppose the probabilism that Joyce wants to justify. Leitgeb and Pettigrew investigate not only whether probabilism and standard conditioning follow from gradational accuracy based on the geometry of reason, but also uniform distribution (their term for the claim of objective Bayesians that there is some principle of indifference for ignorance priors) and Jeffrey conditioning. They show that uniform distribution requires additional axioms which are much less plausible than the ones on the basis of which they derive probabilism and standard conditioning (see Leitgeb and Pettigrew, 2010b, 250f); and that Jeffrey conditioning does not fulfill Joyce's Norm of Gradational Accuracy (see Joyce, 1998, 579) and therefore violates the pursuit of epistemic virtue. 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 One for an amujus, an 'alternative method of updating for Jeffrey-type updating scenarios.' This is List One:

- 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.
- CONFIRMATION An amujus ought to align with intuitions we have about degrees of confirmation.
- 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.

This is grammatically odd, but I'm not sure how you pluralise "amujus"! Maybe "… are both amujuses"?

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 Two, 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. The absolute value is added as a feature to accommodate degree of confirmation functions in subsection 4.7, which may be negative.

- 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.

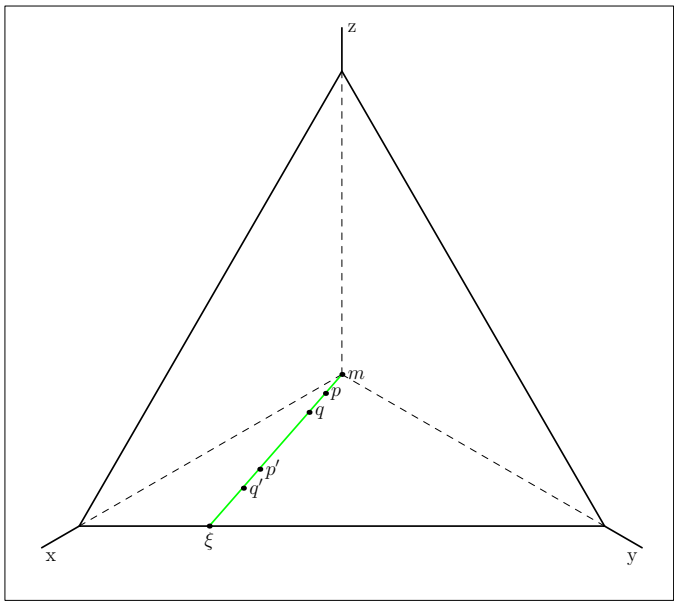

Figure 3: An illustrations of conditions (i)–(iii) for COLLINEAR HORIZON in List Two. p, p' and q, q' must be equidistant and collinear with m and ξ. If q, q' is more peripheral than p, p', then COLLINEAR HORIZON requires that $|d(p, p')| < |d(q, q')|$.

While the Kullback-Leibler divergence of information theory fulfills all the expectations of List One, save HORIZON, it fails all the expectations in List Two. Obversely, the Euclidean distance of the geometry of reason fulfills all the expectations of List Two, save COLLINEAR HORIZON, and fails all the expectations in List One. 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 Two. 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 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 One (see page 10). Consider the following three points in three-dimensional space:

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

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 $D_{\mathrm{KL}}(B, A)$ the Kullback-Leibler divergence of B from A defined as follows, where a_i are the Cartesian coordinates of a:

$$D_{\mathrm{KL}}(B, A) = \sum_{i=1}^{3} b_i \ln \frac{b_i}{a_i}. \tag{13}$$

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 $\|B - A\|$ is defined as in equation (9). What is remarkable about the three points in (12) is that

$$\|C - A\| \approx 0.204 < \|B - A\| \approx 0.212 \tag{14}$$

and

$$D_{\mathrm{KL}}(B, A) \approx 0.0589 < D_{\mathrm{KL}}(C, A) \approx 0.069. \tag{15}$$

The Kullback-Leibler divergence and Euclidean distance give different recommendations with respect to proximity. Assuming the global inaccuracy measure presented in (8) and $E = W$ (all possible worlds are epistemically accessible),

$$\mathrm{GExp}_A(C) \approx 0.653 < \mathrm{GExp}_A(B) \approx 0.656. \tag{16}$$

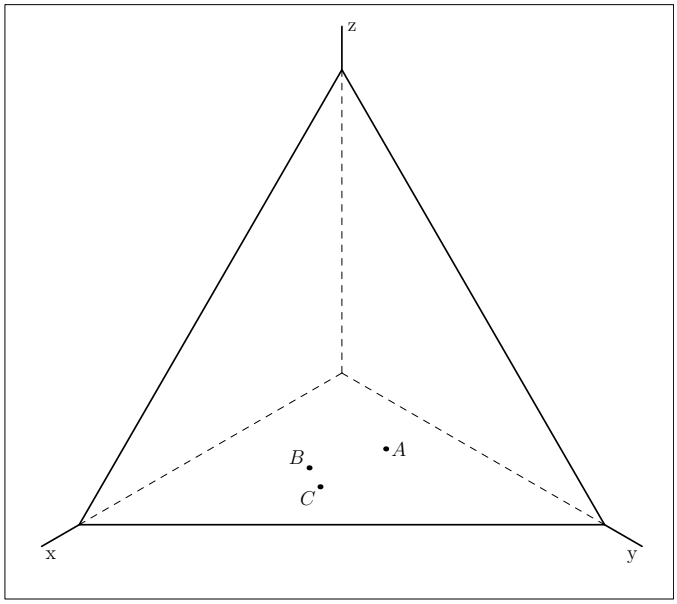

Figure 4: The simplex $\mathbb{S}^2$ in three-dimensional space $\mathbb{R}^3$ with points a, b, c as in equation (12) representing probability distributions A, B, C. Note that geometrically speaking C is closer to A than B is. Using the Kullback-Leibler divergence, however, B is closer to A than C is.

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. There are several things going on here that need some explanation.

3.1 Evaluating Partial Beliefs in Light of Others

We note that for Leitgeb and Pettigrew, expected global inaccuracy of b' is always evaluated by the lights of another partial belief distribution b. This may sound counterintuitive. Should we not evaluate b' by its own lights? It is part of a larger Bayesian commitment that partial belief distributions are not created ex nihilo. They can also not be evaluated for inaccuracy ex nihilo. Leitgeb and Pettigrew say very little about this, but it appears that there is a deeper problem here with the flow of diachronic updating. The classic Bayesian picture is one of moving from a relatively prior probability distribution to a posterior distribution (distinguish relatively prior probability distributions, which precede posterior probability distributions in updating, from absolutely prior probability distributions, which are ignorance priors in the sense that they are not the resulting posteriors of previous updating). This is nicely captured by standard conditioning, Bayes' formula, and updating on the basis of information theory (Jeffrey conditioning, principle of maximum entropy).

The geometry of reason and notions of accuracy based on it sit uncomfortably with this idea of flow, as the suggestion is that partial belief distributions are evaluated on their accuracy without reference to prior probability distributions—why should the accuracy or epistemic virtue of a posterior probability distribution depend on a prior probability distribution which has already been debunked by the evidence? I agree with Leitgeb and Pettigrew that there is no alternative here but to evaluate the posterior by the lights of the prior. Not doing so would saddle us with Carnap's Straight Rule, where priors are dismissed as irrelevant (see Carnap, 1952, 40ff). Yet we shall note that a justification of evaluating a belief function's accuracy by the lights of another belief function is a lot less persuasive than the way Bayesians and information theory integrate prior distributions into forming posterior distributions by virtue of an asymmetric flow of information (see also Shogenji, 2012, who makes a strong case for the influence of prior

probabilities on epistemic justification).

3.2 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

$$\begin{array}{rcl} P(E_1) & = & 1/3 \\ P(E_2) & = & 1/2 \\ P(E_3) & = & 1/6 \end{array} \tag{17}$$

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,

$$\begin{aligned} P'_{\mathrm{JC}}(E_i) &= P'(E_i|F^*)P'(F^*) + P'(E_i|F^{**})P'(F^{**}) \\ &= P(E_i|F^*)P'(F^*) + P(E_i|F^{**})P'(F^{**}) \end{aligned} \tag{18}$$

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 we have called 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:

$$\begin{array}{rcl} P(E_1) + x & = & P'(F^*) \\ P(E_2) + y + P(E_3) + y & = & P'(F^{**}) \end{array} \tag{19}$$

and then set

$$\begin{array}{rcl} P'_{\mathrm{LP}}(E_1) & = & P(E_1) + x \\ P'_{\mathrm{LP}}(E_2) & = & P(E_2) + y \\ P'_{\mathrm{LP}}(E_3) & = & P(E_3) + y \end{array} \tag{20}$$

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

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

Compare these results to the results of Jeffrey conditioning:

$$\begin{array}{rcl} P'_{\mathrm{JC}}(E_1) & = & 1/2 \\ P'_{\mathrm{JC}}(E_2) & = & 3/8 \\ P'_{\mathrm{JC}}(E_3) & = & 1/8 \end{array} \tag{22}$$

Note that (17), (22), and (21) correspond to A, B, C in (12).

3.3 Triangulating LP and Jeffrey Conditioning

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

$$d(A, C) = d(A, B) + d(B, C) \tag{23}$$

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 we will address in detail in subsection 5.1). Call B a zero-sum point between A and C if (23) 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 poins 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:

$$D_{\mathrm{KL}}(C, A) = D_{\mathrm{KL}}(B, A) + D_{\mathrm{KL}}(C, B) \tag{24}$$

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

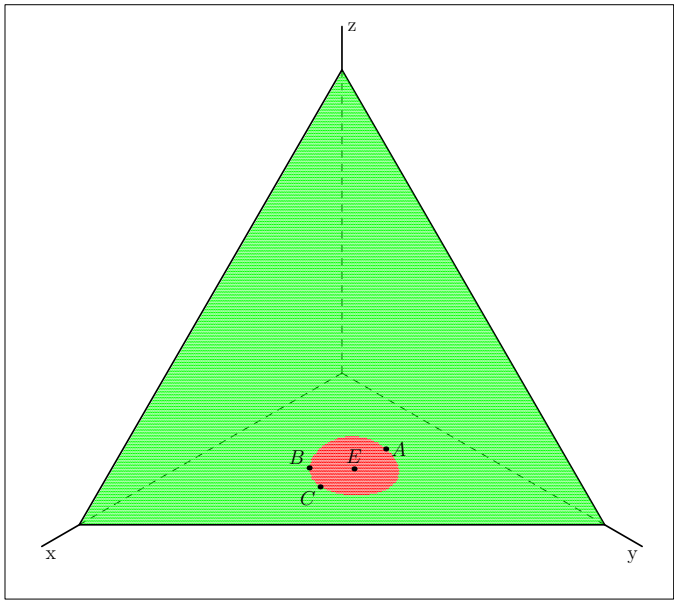

Figure 5: The zero-sum line between A and C is the boundary line between the green area, where the triangle inequality holds, and the red area, where the triangle inequality is violated. The posterior probability distribution B recommended by Jeffrey conditioning always lies on the zero-sum line between the prior A and the LP posterior C, as per equation (24). E is the point in the red area where the triangle inequality is most efficiently violated.

what from the perspective of information theory (!) looks like a detour, i.e. the triangle inequality does not hold.

To prove equation (24) 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:

$$
\begin{aligned}
A &= (1-\alpha, \beta, \alpha-\beta) \\
B &= \left(1-\gamma, \tfrac{\gamma\beta}{\alpha}, \tfrac{\gamma(\alpha-\beta)}{\alpha}\right) \\
C &= \left(1-\gamma, \beta+\tfrac{1}{2}(\gamma-\alpha), \alpha-\beta+\tfrac{1}{2}(\gamma-\alpha)\right)
\end{aligned} \tag{25}
$$

The rest is basic algebra using the definition of the Kullback-Leibler divergence in (13). To prove the claim for arbitrary n one simply generalizes (25). It is a handy corollary of (24) that whenever (A, B) and (B, C) violate TRANSITIVITY OF ASYMMETRY then

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

in violation of TRIANGULARITY. This way we will not have to go hunting for an example to demonstrate the violation of TRIANGULARITY. A, B, C of (12) fulfill all the conditions for (26) and therefore violate TRIANGULARITY.

It is an interesting question to wonder which point E violates the triangle inequality most efficiently so that

$$
D_{\mathrm{KL}}(E, C) + D_{\mathrm{KL}}(A, E) \tag{27}
$$

is minimal. Let $e = (e_1, \ldots, e_n)$ represent E in $\mathbb{S}^{n-1}$. Use the Lagrange Multiplier method to find the Lagrangian

$$
\mathcal{L}(e, \lambda) = \sum_{i=1}^{n} e_i \log \frac{e_i}{a_i} + \sum_{i=1}^{n} c_i \log \frac{c_i}{e_i} + \lambda \left(\sum_{i=1}^{n} e_i - 1 \right) \tag{28}
$$

The Lagrange Multiplier method gives us

$$
\frac{\partial \mathcal{L}}{\partial e_k} = \log \frac{e_k}{a_k} + 1 - \frac{r}{q} + \lambda = 0 \text{ for each } k = 1, \ldots, n. \tag{29}
$$

Manipulate this equation to yield

$$\frac{c_k}{e_k} \exp\left(\frac{c_k}{e_k}\right) = \frac{c_k}{a_k} \exp(1 + \lambda). \tag{30}$$

To solve (30), use the Lambert W function

$$e_k = \frac{c_k}{W\left(\frac{c_k}{a_k} \exp(1 + \lambda)\right)}. \tag{31}$$

Choose λ to fulfill the constraint $\sum e_i = 1$. The result for the discrete case accords with Ovidiu Calin and Constantin Udriste's result for the continuous case (see equation 4.7.9 in Calin and Udriste, 2014, 127). Numerically, for A and C as defined in equation (12),

$$E = (0.415, 0.462, 0.123). \tag{32}$$

This is subtly different from the midpoint $m_i = 0.5a_i + 0.5c_i$ (if we were minimizing $D_{\mathrm{KL}}(A, E) + D_{\mathrm{KL}}(C, E)$, the solution would be the midpoint). I do not know whether A, E, C are collinear (see figure 5 for illustration).

4 Expectations for the Geometry of Reason

This section provides more detail for the expectations in List One (see page 10) 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

$$\begin{array}{lcl} P'_{\text{SC}}(E_1) & = & 0 \\ P'_{\text{SC}}(E_2) & = & 3/4 \\ P'_{\text{SC}}(E_3) & = & 1/4. \end{array} \tag{33}$$

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

$$\begin{array}{lcccl} P'_{\text{JC}}(E_1) & = & 1/n & \to & 0 \\ P'_{\text{JC}}(E_2) & = & 3(n-1)/4n & \to & 3/4 \\ P'_{\text{JC}}(E_3) & = & (n-1)/4n & \to & 1/4, \end{array} \tag{34}$$

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

$$\begin{array}{lcccl} P'_{\text{LP}}(E_1) & = & 1/n & \to & 0 \\ P'_{\text{LP}}(E_2) & = & (4n-3)/6n & \to & 2/3 \\ P'_{\text{LP}}(E_3) & = & (2n-5)/6n & \to & 1/3. \end{array} \tag{35}$$

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 (17).

Then his posterior probabilities look as follows:

$$\begin{array}{rcl} P'_{\text{JC}}(E_1) & = & 2/3 \\ P'_{\text{JC}}(E_2) & = & 1/4 \\ P'_{\text{JC}}(E_3) & = & 1/12 \end{array} \tag{36}$$

$$\begin{array}{rcl} P'_{\text{LP}}(E_1) & = & 2/3 \\ P'_{\text{LP}}(E_2) & = & 1/3 \\ P'_{\text{LP}}(E_3) & = & 0 \end{array} \tag{37}$$

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 (we 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

$$\begin{array}{rcl} Q(E_1) & = & 1/3 \\ Q(E_2^*) & = & 1/4 \\ Q(E_2^{**}) & = & 1/4 \\ Q(E_3) & = & 1/6. \end{array} \tag{38}$$

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,

$$\begin{array}{rcl} P'_{\text{JC}}(E_1) & = & 1/2 \\ P'_{\text{JC}}(E_2) & = & 3/8 \\ P'_{\text{JC}}(E_3) & = & 1/8 \end{array} \tag{39}$$

$$\begin{array}{rcl} Q'_{\text{JC}}(E_1) & = & 1/2 \\ Q'_{\text{JC}}(E_2^*) & = & 3/16 \\ Q'_{\text{JC}}(E_2^{**}) & = & 3/16 \\ Q'_{\text{JC}}(E_3) & = & 1/8 \end{array} \tag{40}$$

they do not agree if they use LP conditioning,

$$\begin{array}{rcl} P'_{\text{LP}}(E_1) & = & 1/2 \\ P'_{\text{LP}}(E_2) & = & 5/12 \\ P'_{\text{LP}}(E_3) & = & 1/12 \end{array} \tag{41}$$

$$\begin{array}{rcl} Q'_{\text{LP}}(E_1) & = & 1/2 \\ Q'_{\text{LP}}(E_2^*) & = & 7/36 \\ Q'_{\text{LP}}(E_2^{**}) & = & 7/36 \\ Q'_{\text{LP}}(E_3) & = & 1/9. \end{array} \tag{42}$$

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

$$\begin{array}{rcl} P(X_1) & = & 0 \\ P(X_2) & = & 0.3 \\ P(X_3) & = & 0.6 \\ P(X_4) & = & 0.1 \end{array} \tag{43}$$

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 (43) or on the following list:

$$\begin{array}{rcl} P(X_2) & = & 0.3 \\ P(X_3) & = & 0.6 \\ P(X_4) & = & 0.1 \end{array} \tag{44}$$

Whether you update on (43) or (44) 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

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)

For an interesting instance of the horizon effect in asymmetric multi-dimensional scaling see Chino and Shiraiwa, 1993, section 3, where Naohito Chino and Kenichi Shiraiwa describe as one of the properties of their Hilbert space models of asymmetry how "the similarity between the pair of objects located far from the centroid of objects, say, the origin, is greater than that located near the origin, even if their distances are the same" (42).

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 Two (see page 11). 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.

In the next section, I will closely tie the issue of the horizon effect to confirmation. The two main candidates for quantitative measures of relevance confirmation disagree on precisely this issue. Whether you, the reader, will accept the horizon requirement may depend on what your view is on degree of confirmation theory.

4.7 Confirmation

The geometry of reason thinks about the comparison of probability distributions in terms of distance. Information theory thinks about the comparison along the lines of information loss when one distribution is used to encode a message rather than the other distribution. One way to test these approaches is to ask how well they align with a third approach to such a comparison: degree of confirmation. Our main concern is the horizon effect of the previous subsection. Which approaches to degree of confirmation theory reflect it, and how do these approaches correspond to the disagreements between information theory and the geometry of reason?

There is, of course, a relevant difference between the aims of the epistemic utility approach to updating and the aims of degree of confirmation theory. The former investigates norms to which a rational agent conforms in her pursuit of epistemic virtue. The latter seeks to establish qualitative and quantitative measures of impact that evidence has on a hypothesis. Both, however, (I will restrict my attention here to quantitative degree of confirmation theory) attend to the probability of an event, which degree of confirmation theory customarily calls h for hypothesis, before and after the rational agent processes another event, customarily called e for evidence, i.e. $x = P(h|k)$ and $y = P(h|e, k)$ (k is background information).

For perspectives on the link between confirmation and information see Shogenji, 2012, 37f; Crupi and Tentori, 2014; and Milne, 2014, section 4. Vincenzo Crupi and Katya Tentori suggest that there is a "parallelism between confirmation and information search [which] can serve as a valuable heuristic for theoretical work" (Crupi and Tentori, 2014, 89).

In degree of confirmation theory, incremental confirmation is distinguished from absolute confirmation in the following sense. Let h be the presence of a very rare disease and e a test result such that $y >> x$ but $y < 1 - y$. Then, absolutely speaking, e disconfirms h (for Rudolf Carnap, absolute confirmation involves sending y above a threshold r which must be greater than or

equal to 0.5). Absolute confirmation is not the subject of this section. I will exclusively discuss incremental confirmation (also called relevance confirmation, just as absolute confirmation is sometimes called firmness confirmation) where $y > x$ implies (incremental) confirmation, $y < x$ implies (incremental) disconfirmation, and $y = x$ implies the lack of both. The difference is illustrated in figure 6.

All proposed measures of quantitative, incremental degree of confirmation considered here are a function of x and y. Dependence of incremental confirmation on only x and y is not trivial, as $P(e|k)$ and $P(e|h,k)$ cannot be expressed using only x and y (for a case why dependence should be on only x and y see Atkinson, 2012, 50, with an irrelevant conjunction argument; and Milne, 2014, 254, with a continuity argument). David Christensen's measure $P(h|e,k) - P(h|\neg e,k)$ (see Christensen, 1999, 449) and Robert Nozick's $P(e|h,k) - P(e|\neg h,k)$ (see Nozick, 1981, 252) are not only dependent on x and y, but also on $P(e|k)$, which makes them vulnerable to Atkinson's and Milne's worries just cited.

Consider the following six contenders for a quantitative, incremental degree of confirmation function, dependent on only x and y. They are based on, in a brief slogan, (i) difference of conditional probabilities, (ii) ratio of conditional probabilities, (iii) difference of odds, (iv) ratio of likelihoods, (v) Gaifman's treatment of Hempel's raven paradox, and (vi) conservation of contrapositivity and commutativity. Logarithms throughout this paper are assumed to be the natural logarithm in order to facilitate easy differentiation, although generally a particular choice of base (greater than one) does not make a relevant difference.

$$
\begin{aligned}
&\text{(i)} && M_P(x,y) &&= y - x \\
&\text{(ii)} && R_P(x,y) &&= \log \frac{y}{x} \\
&\text{(iii)} && J_P(x,y) &&= \frac{y}{1-y} - \frac{x}{1-x} \\
&\text{(iv)} && L_P(x,y) &&= \log \frac{y(1-x)}{x(1-y)} \\
&\text{(v)} && G_P(x,y) &&= \log \frac{1-x}{1-y} \\
&\text{(vi)} && Z_P(x,y) &&= \begin{cases} \dfrac{y-x}{1-x} & \text{if } y \geq x \\ \dfrac{y-x}{x} & \text{if } y < x \end{cases}
\end{aligned}
\tag{45}
$$

M_P is defended by Carnap, 1962; Earman, 1992; Rosenkrantz, 1994. R_P is defended by Keynes, 1921; Milne, 1996; Shogenji, 2012. J_P is defended by Festa, 1999. L_P is defended by Good, 1950; Good, 1983, chapter 14; Fitelson, 2006; Zalabardo, 2009. G_P is defended by Gaifman, 1979, 120, without the logarithm (I added it to make G_P more comparable to the other functions). Z_P is defended by Crupi et al., 2007. For more literature supporting the various measures consult footnote 1 in Fitelson, 2001, S124; and an older survey of options in Kyburg, 1983.

Footnote?

To compare how these degree of confirmation measures align with the concept of difference between probability distributions for the purpose of updating it is best to look at derivatives as they reflect the rate of change from the middle to the extremes. This is how we capture the horizon effect requirement for two dimensions. One important difference between degree of confirmation theory and updating is that the former is concerned with a hypothesis and its negation whereas the latter considers all sorts of domains for the probability distribution (in this paper, I have restricted myself to a finite outcome space). As far as the analogy between degree of confirmation theory on the one hand and updating on the other hand is concerned, we only need to look at the two-dimensional case.

To discriminate between candidates (i)–(vi), I am setting up three criteria

(complementing many others in the literature). Let $D(x, y)$ be the generic expression for the degree of confirmation function. Call this List Three.

- ADDITIVITY A theory can be confirmed piecemeal. Whether the evidence is split up into two or more components or left in one piece is irrelevant to the amount of confirmation it confers. Formally, $D(x, z) = D(x, y) + D(y, z)$. Note that this is not the usual triangle inequality because we are in two dimensions.

- SKEW-ANTISYMMETRY It does not matter whether h or $\neg h$ is in view. Confirmation and disconfirmation are commensurable. Formally, $D(x, y) = -D(1 - x, 1 - y)$. A surprising number of candidates fail this requirement, and the requirement is not common in the literature (see, however, the second clause in Milne's fourth desideratum in 1996, 21). In defence of this requirement consider example 8 below. $d_1 > d_2$ may have a negative impact on the latter scientist's grant application, even though the inequality may solely be due to a failure to fulfill skew-antisymmetry.

- CONFIRMATION HORIZON An account of degree of confirmation must exhibit the horizon effect as in List One and List Two, except more simply in two dimensions. Formally, the functions $\partial D^+_\varepsilon/\partial x$ must be strictly positive and the functions $\partial D^-_\varepsilon/\partial x$ must be strictly negative for all $\varepsilon \in (-1/2, 1/2)$. These functions are defined in (46) and (47), and I prove in appendix C that the requirement to keep them strictly positive/negative is equivalent to the horizon effect as described formally in List Two (see page 11).

Example 8: Grant Adjudication I. Two scientists compete for grant money. Professor X presents an experiment conferring degree of confirmation d_1 on a hypothesis, if successful; Professor Y presents an experiment conferring degree of disconfirmation $-d_2$ on the negation of the same hypothesis, if unsuccessful. (For the relevance of quantitative confirmation measures to the evaluation of scientific projects see Salmon, 1975, 11.)

The functions for the horizon effect are defined as follows. Let $\varepsilon \in (-1/2, 1/2)$ be fixed. Recall that $D(x, y)$ is the generic expression for a confirmation function measuring the degree of confirmation that a posterior $y = P(h|e, k)$ bestows on a hypothesis for which the prior is $x = P(h|k)$. ε is the difference $y - x$. For $\varepsilon > 0$,

$$\begin{array}{lll} D^-_\varepsilon : (0, \frac{1}{2} - \varepsilon) \to \mathbb{R} & D^-_\varepsilon(x) & = & |D(x, x+\varepsilon)| \\ D^+_\varepsilon : (\frac{1}{2}, 1 - \varepsilon) \to \mathbb{R} & D^+_\varepsilon(x) & = & |D(x, x+\varepsilon)| \end{array} \tag{46}$$

For $\varepsilon < 0$,

$$\begin{array}{lll} D^-_\varepsilon : (-\varepsilon, \frac{1}{2}) \to \mathbb{R} & D^-_\varepsilon(x) & = & |D(x, x+\varepsilon)| \\ D^+_\varepsilon : (\frac{1}{2} - \varepsilon, 1) \to \mathbb{R} & D^+_\varepsilon(x) & = & |D(x, x+\varepsilon)| \end{array} \tag{47}$$

The rate of change for the different quantitative measures of degree of confirmation can be observed in figure 6. The pass and fail verdicts in the table below are evident from figure 6 and the table of derivatives provided in appendix C. Only J_P, L_P and Z_P fulfill the horizon requirement.

Candidate	*Triangularity*	*Skew-Antisymmetry*	*Confirmation Horizon*
M_P	pass	pass	fail
R_P	pass	fail	fail
J_P	pass	fail	pass
L_P	pass	pass	pass
G_P	pass	fail	fail
Z_P	fail	pass	pass

The table makes clear that only L_P passes all three tests. I am not making a strong independent case for L_P here, especially against M_P, which is the most likely hero of the geometry of reason. This has been done elsewhere (for example in Schlesinger, 1995, where L_P and M_P are compared to each other in their performance given some intuitive examples; Elliott Sober presents the counterargument in Sober, 1994). The argumentative force of this subsection appeals to those who are already sympathetic to L_P. Adherents of M_P will hopefully find other items on List One (see page 10) persuasive and reject the geometry of reason, in which case they may come back to this subsection and re-evaluate their commitment to M_P.

Example 9: Grant Adjudication II. Two scientists compete for grant money. Professor X presents an experiment that will increase the probability of a hypothesis from 98% to 99%, if successful. Professor Y presents an experiment that will increase the probability of a hypothesis from 1% to 2%, if successful.

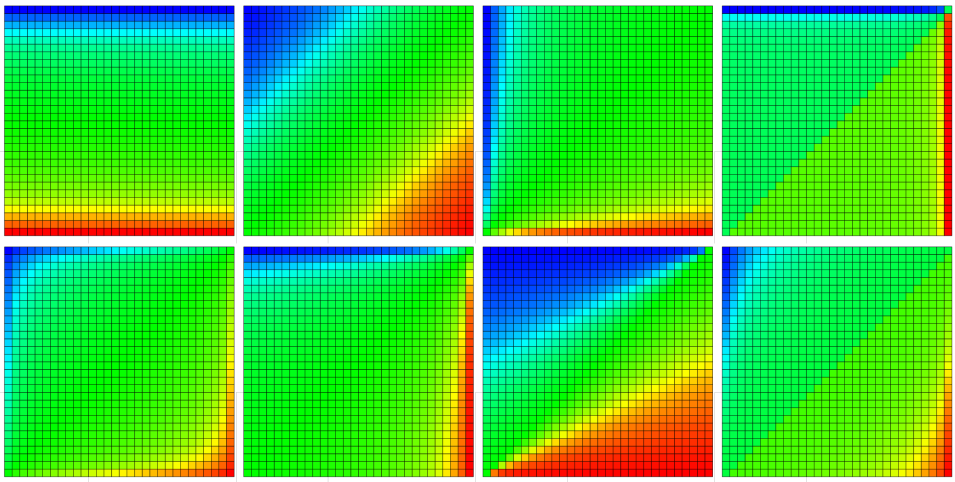

Figure 6: Illustration for the six degree of confirmation candidates plus Carnap's firmness confirmation and the Kullback-Leibler divergence. The top row, from left to right, illustrates FMRJ, the bottom row LGZI. 'F' stands for Carnap's firmness confirmation measure $F_P(x, y) = \log(y/(1-y))$. 'M' stands for candidate (i), $M_P(x, y)$ in (45), the other letters correspond to the other candidates (ii)-(v). 'I' stands for the Kullback-Leibler divergence multiplied by the sign function of $y - x$ to mimic a quantitative measure of confirmation. For all the squares, the colour reflects the degree of confirmation with x on the x-axis and y on the y-axis, all between 0 and 1. The origin of the square's coordinate system is in the bottom left corner. Blue signifies strong confirmation, red signifies strong disconfirmation, and green signifies the scale between them. Perfect green is $x = y$. G_P looks like it might pass the horizon requirement, but the derivative reveals that it fails CONFIRMATION HORIZON (see appendix C).

All else being equal, Professor Y should receive the grant money. If her experiment is more successful, it will arguably make more of a difference. This example illustrates that the analogy between degree of confirmation and updating remains tenuous, since for degree of confirmation theory the consensus on intuitions is far inferior to the updating case. If, for example, the confirmation function is anti-symmetric and $D(x, y) - D(y, x)$ is zero (for M_P and L_P, for example), then together with skew-antisymmetry this means that degree of confirmation is equal for Professor X and Professor Y. Despite its three passes in the above table, L_P fails here.

I didn't get the intuition here, FWIW. Maybe explain/argue?

Based on Roberto Festa's J_P, Professor X's prospective degree of confirmation is 5000 times larger than Professor Y's, but Festa in particular insists "that there is no universally applicable P-incremental c-measure, and that the appropriateness of a P-incremental c-measure is highly context-dependent" (Festa, 1999, 67). R_P and Z_P appear to be sensitive to example

9. The Kullback-Leibler divergence gives us the right result as well, where the degree of confirmation for going from 1% to 2% is $3.91 \cdot 10^{-3}$ compared to $3.12 \cdot 10^{-3}$ for going from 98% to 99%, but the Kullback-Leibler divergence is not a serious degree of confirmation candidate. It fulfills SKEW-ANTISYMMETRY and CONFIRMATION HORIZON, but not ADDITIVITY (see subsection 5.1).

Intuitions easily diverge here. Christensen may be correct when he says, "perhaps the controversy between difference and ratio-based positive relevance models of quantitative confirmation reflects a natural indeterminateness in the basic notion of 'how much' one thing supports another" (Christensen, 1999, 460). Pluralists allow therefore for "distinct, complementary notions of evidential support" (Hájek and Joyce, 2008, 123). I am sympathetic towards this indeterminateness in degree of confirmation theory, but not when it comes to updating (see the full employment theorem in Lukits, 2013, 1413).

This subsection assumes that despite these problems with the strength of the analogy, degree of confirmation and updating are sufficiently similar to be helpful in associating options with each other and letting the arguments in each other's favour and disfavour cross-pollinate. As an aside, Christensen's S-support given by evidence E is stable over Jeffrey conditioning on $[E, \neg E]$; LP-conditioning is not (see Christensen, 1999, 451). This may serve as another argument from degree of confirmation theory in favour of information theory (which supports Jeffrey conditioning) against the geometry of reason (which supports LP conditioning).

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

As mentioned at the end of subsection 3.3, the three points A, B, C in (12) violate TRIANGULARITY as in (26):

$$D_{\mathrm{KL}}(A, C) > D_{\mathrm{KL}}(B, C) + D_{\mathrm{KL}}(A, B). \tag{48}$$

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 $\delta(x, z) = D_{\mathrm{KL}}(z, x)$. 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:

$$\delta(x, z) > \delta(x, y) + \delta(y, z). \tag{49}$$

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 you defining these partitions, but note that any dividing point $y^{j_0 k}$ will also be a dividing point in the more finely grained partitions y^{jk} with $j \geq j_0$. Then define the sequence

$$T_j = \sum_k \delta\left(y^{jk}, y^{j(k+1)}\right) \tag{50}$$

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 (49), 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 (49), it is straightforward to see that (49) is equivalent to

$$\sum_{i=1}^{n}(z_i - x_i)\log\frac{\vartheta x_i + (1-\vartheta)z_i}{x_i} > 0. \tag{51}$$

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

$$\sum_{i=1}^{n}\left(z_i + \frac{\vartheta}{1-\vartheta}x_i - \frac{\vartheta}{1-\vartheta}x_i - x_i\right)\log\frac{\frac{1}{1-\vartheta}\left(\vartheta x_i + (1-\vartheta)z_i\right)}{\frac{1}{1-\vartheta}x_i} > 0. \tag{52}$$

(52) is clearly equivalent to (51). It is also equivalent to

$$\sum_{i=1}^{n}\left(z_i + \frac{\vartheta}{1-\vartheta}x_i\right)\log\frac{z_i + \frac{\vartheta}{1-\vartheta}x_i}{\frac{1}{1-\vartheta}x_i} + \sum_{i=1}^{n}\frac{1}{1-\vartheta}x_i\log\frac{\frac{1}{1-\vartheta}x_i}{z_i + \frac{\vartheta}{1-\vartheta}x_i} > 0, \tag{53}$$

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 those adduced by Schlesinger 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 Two 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 we 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.

$$p = \left(\frac{1}{5}, \frac{2}{5}, \frac{2}{5}\right) \qquad p' = q = \left(\frac{1}{4}, \frac{3}{8}, \frac{3}{8}\right) \qquad q' = \left(\frac{3}{10}, \frac{7}{20}, \frac{7}{20}\right) \tag{54}$$

The conditions of COLLINEAR HORIZON in List Two (see page 11) 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,

$$D_{\mathrm{KL}}(B, A) = 7.3820 \cdot 10^{-3} > 6.4015 \cdot 10^{-3} = D_{\mathrm{KL}}(C, B). \tag{55}$$

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 7 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 $x_i = 0.5(m_i + y_i)$.

$$\text{(i) } D_{\mathrm{KL}}(Y, M) > D_{\mathrm{KL}}(M, Y) \text{ and (ii) } D_{\mathrm{KL}}(X, M) > D_{\mathrm{KL}}(Y, X) \tag{56}$$

In terms of coordinates, the inequalities reduce to

$$\text{(i) } H(y) < \frac{1}{n}\sum(\log y_i) - \log\frac{1}{n^2} \text{ and} \tag{57}$$

$$\text{(ii) } H(y) > \log\frac{4}{n} - \sum\left[\left(\frac{3}{2}y_i + \frac{1}{2n}\right)\log\left(y_i + \frac{1}{n}\right)\right]. \tag{58}$$

(i) is simply the case described in the next subsection for asymmetry and illustrated on the bottom left of figure 8. (ii) tells us how far from the midpoint we can go with a scenario where $p = m, p' = q$ while violating COLLINEAR HORIZON. Clearly, as illustrated in figure 7, there is a relationship between asymmetry and COLLINEAR HORIZON.

Phrasing here is not stylistically in keeping with the rest of the paper.

The bitter aftertaste that remains with COLLINEAR HORIZON is that 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.

5.3 Transitivity of Asymmetry

Recall Joyce's two axioms Weak Convexity and Symmetry (see page 9). The geometry of reason (certainly in its Euclidean form) mandates Weak Convexity because the bisector of an isosceles triangle is always shorter

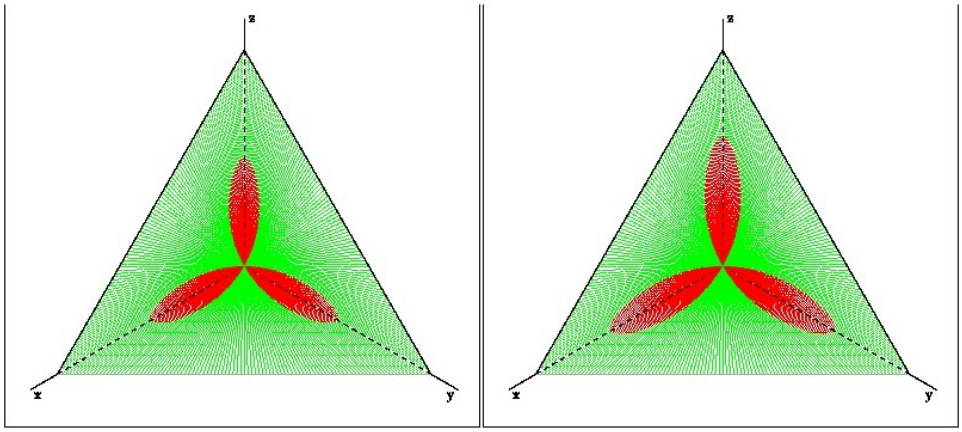

Figure 7: These two diagrams illustrate inequalities (57) and (58). The former displays all points in red which violate COLLINEAR HORIZON, measured from the centre. The latter displays points in different colours whose orientation of asymmetry differs, measured from the centre. The two red sets are not the same, but there appears to be a relationship, one that ultimately I suspect to be due to the more basic property of asymmetry.

than the isosceles sides. Weak Convexity, on the one hand, also holds for information theory (see appendix A for a proof). Symmetry, on the other hand, 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 10: 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

$$\Delta_P(P') = D_{\mathrm{KL}}(P', P) - D_{\mathrm{KL}}(P, P'), \tag{59}$$

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

$$\Delta^{-1}(\mathbb{R}_{>0}) \qquad \Delta^{-1}(\mathbb{R}_{<0}) \qquad \Delta^{-1}(\{0\}) \tag{60}$$

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 (see appendix B).

In higher-dimensional cases, however, the tripartite partition (60) 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 8 for graphical illustration of this point.

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 Two on page 11. For an example of it, consider

$$P_1 = \left(\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\right) \qquad P_2 = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right) \qquad P_3 = \left(\frac{2}{5}, \frac{2}{5}, \frac{1}{5}\right) \quad (61)$$

In the terminology of TRANSITIVITY OF ASYMMETRY in List Two, (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 (61) 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.

Footnote?

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. I am explaining the reasons in appendix D.

The failure of Chino's modeling approach to make useful distinctions among asymmetric distance measures between probability distributions leads us to the more complex theory of information geometry and differentiable manifolds. Both the results of Shun-ichi Amari (see Amari, 1985; and Amari and Nagaoka, 2000) and Nikolai Chentsov (see Chentsov, 1982) serve to highlight the special properties of the Kullback-Leibler divergence, not without elevating the discussion to a level of mathematical sophistication, however, where it is difficult to retain the appeal to epistemic intuitions. Information geometry considers probability distributions as differentiable manifolds equipped with a Riemannian metric. This metric, however, is Fisher's information metric, not the Kullback-Leibler divergence, and it is defined on the tangent space of the simplex representing finite-dimensional probabil-

ity distributions. There is a sense in which the Fisher information metric is the derivative of the Kullback-Leibler divergence, and so the connection to epistemic intuitions can be re-established.

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 One, 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 Two—expectations which the geometry of reason fulfills. We are 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.

A Appendix: Weak Convexity and Symmetry in Information Geometry

Using information theory instead of the geometry of reason, Joyce's result still stands, vindicating probabilism on epistemic merits rather than prudential ones: partial beliefs which violate probabilism are dominated by partial beliefs which obey it, no matter what the facts are.

Joyce's axioms, however, will need to be reformulated to accommodate asymmetry. This appendix shows that the axiom Weak Convexity (see section 2) still holds in information geometry. Consider three points $Q, R, S \in \mathbb{S}^{n-1}$ (replace $\mathbb{S}^{n-1}$ by the n-dimensional space of non-negative real numbers, if you do not want to assume probabilism) for which

$$D_{\mathrm{KL}}(Q, R) = D_{\mathrm{KL}}(Q, S). \tag{62}$$

I will show something slightly stronger than Weak Convexity: Joyce's inequality is not only true for the midpoint between R and S but for all points $\vartheta R + (1 - \vartheta)S$, as long as $0 \leq \vartheta \leq 1$. The inequality aimed for is

$$D_{\mathrm{KL}}(Q, \vartheta R + (1 - \vartheta)S) \leq D_{\mathrm{KL}}(Q, R) = D_{\mathrm{KL}}(Q, S). \tag{63}$$

To show that it holds I need the log-sum inequality, which is a result of Jensen's inequality (for a proof of the log-sum inequality see Theorem 2.7.1 in Cover and Thomas, 2006, 31). For non-negative numbers $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$,

$$\sum_{i=1}^{n} a_i \ln \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} a_i \right) \ln \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}. \tag{64}$$

(63) follows from (64) via

$$D_{\mathrm{KL}}(Q,R) = \vartheta D_{\mathrm{KL}}(Q,R) + (1-\vartheta) D_{\mathrm{KL}}(Q,S) =$$
$$\sum_{i=1}^{n} \left(\vartheta q_i \ln \frac{\vartheta q_i}{\vartheta r_i} + (1-\vartheta) q_i \ln \frac{(1-\vartheta) q_i}{(1-\vartheta) s_i} \right) \geq$$
$$\sum_{i=1}^{n} q_i \ln \frac{q_i}{\vartheta r_i + (1-\vartheta) s_i} = D_{\mathrm{KL}}(Q, \vartheta R + (1-\vartheta) S). \tag{65}$$

I owe some thanks to physicist friend Thomas Buchberger for help with this proof. Interested readers can find a more general claim in Csiszár's Lemma 4.1 (see Csiszár and Shields, 2004, 448), which accommodates convexity of the Kullback-Leibler divergence as a special case.

B Appendix: Asymmetry in Two Dimensions

This appendix contains a proof that the threefold partition (60) of $\mathbb{S}^1$ is well-behaved, in contrast to the threefold partition of $\mathbb{S}^2$ as illustrated by figure 8. For the two-dimensional case, i.e. considering $p, q \in \mathbb{S}^1$ with $0 < p, q < 1, p + p' = 1$ and $q + q' = 1$,

$$\begin{array}{lllllll} \Delta_q(p) & > & 0 & \text{for} & |p-p'| & > & |q-q'| \\ \Delta_q(p) & = & 0 & \text{for} & |p-p'| & = & |q-q'| \\ \Delta_q(p) & < & 0 & \text{for} & |p-p'| & < & |q-q'| \end{array} \tag{66}$$

where $\Delta_q(p) = D_{\mathrm{KL}}(q,p) - D_{\mathrm{KL}}(p,q)$ and $D_{\mathrm{KL}}(p,q) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}$. Part of information theory's ill behaviour outlined in section 5.3 is that in the higher-dimensional case the partition does not follow the simple rule that higher entropy of P compared to Q implies that $\Delta_Q(P) > 0$ (Δ here defined as in (59)). In the two-dimensional case, however, this simple rule applies.

That a comparison in entropy $H(p) = -p \log p - (1-p) \log(1-p)$ between $H(p)$ and $H(q)$ corresponds to a comparison of $|p-p'|$ and $|q-q'|$ is trivial. The proof for (66) is straightforward given the following non-trivial lemma establishing a very tight inequality. Given that $p + p' = 1$ and $q + q' = 1$ and $p, q, p', q' > 0$ it is true that

$$\text{If } \log(p/q) > \log(q'/p') \text{ then } (p+q)\log(p/q) > (p'+q')\log(q'/p') \quad (67)$$

Let $x = p/q$ and $y = q'/p'$. We know that $x > y$ since $\log x > \log y$. Now we want to show $(p+q)\log x > (p'+q')\log y$. Note that $p = xq, q' = p'y, p+q = q(x+1)$, and $p' + q' = p'(y+1)$. Therefore,

$$q = \frac{1-y}{1-xy} \quad (68)$$

and

$$p' = \frac{1-x}{1-xy}. \quad (69)$$

What we want to show is that $x > y$ implies

$$\frac{1-y}{1-xy}(x+1)\log x > \frac{1-x}{1-xy}\log y. \quad (70)$$

Note that $f(x) = (1-x)^{-1}(x+1)\log x$ is increasing on $(0,1)$ and decreasing on $(1,\infty)$, and consider the following two cases:

(i) When $x < 1, y < 1$, (70) follows from the fact that f is increasing on $(0,1)$.

(ii) When $x > 1, y > 1$, (70) follows from the fact that f is decreasing on $(1,\infty)$.

Mixed cases such as $x > 1, y < 1$ do not occur, as for example $x > 1$ implies $y > 1$.

C Appendix: The Horizon Requirement Formalized

Consider the following two conditions on a difference measure D on a simplex $\mathbb{S}^{n-1} \subset \mathbb{R}^n$, which is assumed to be a smooth function from $\mathbb{S}^{n-1} \times \mathbb{S}^{n-1} \to \mathbb{R}$.

(h1) 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 condition see figure 3.

(h2) Let $\mu \in (-1, 1)$ be fixed and D_μ defined as in (72). Then $dD_\mu/dx > 0$, where dD_μ/dx is the total derivative as x moves towards $\xi(x)$, the unique boundary point which is collinear with x and m.

To define D_μ, the hardest part is to specify the domain. Let this domain $V(\mu) \subseteq \mathbb{S}^{n-1}$ be defined as

$$V(\mu) = \begin{cases} \{x \in \mathbb{S}^{n-1} | x_i < (1-\mu)\xi_i(x) + \mu m_i, i = 1, \ldots, n\} & \text{for } \mu > 0 \\ \{x \in \mathbb{S}^{n-1} | x_i > (1+\mu)m_i - \mu\xi_i(x), i = 1, \ldots, n\} & \text{for } \mu < 0. \end{cases} \tag{71}$$

Then $D_\mu : V(\mu) \to \mathbb{R}_0^+$ is defined as

$$D_\mu(x) = |D(x, y(x))| \tag{72}$$

where $y_i(x) = x_i + \mu(\xi_i(x) - m_i)$. Remember that $\xi(x)$ is the unique boundary point which is collinear with x and m. Now for the proof that (h1) and (h2) are equivalent.

First assume (h1) and the negation of (h2). Since D is smooth, there must be a $\bar{\mu}$ and two points x' and x'' collinear with m and a boundary point $\bar{\xi}$ such that $D_{\bar{\mu}}(x') \geq D_{\bar{\mu}}(x'')$ even though $\|\bar{\xi} - x''\| < \|\bar{\xi} - x'\|$. If this were not the case, D_μ would be strictly increasing running towards the boundary points for all μ and its total derivative would be strictly positive so that (h2) follows. Now consider the four points x', x'', y', y'' where $y_i' = x_i' + \mu(\bar{\xi}_i - m_i)$ and $y_i'' = x_i'' + \mu(\bar{\xi}_i - m_i)$ for $i = 1, \ldots, n$. Without loss of generality, assume $\bar{\mu} > 0$. Then x', x'', y', y'' fulfill the conditions in (h1) and $D_{\bar{\mu}}(x') < D_{\bar{\mu}}(x'')$, in contradiction to the aforesaid.

Then assume (h2). Let x', x'', y', y'' be four points as in (h1). Consider $\mu = \|\xi - m\|/\|x'' - x'\|$. Then $D_\mu(x') = |D(x', x'')|$ and $D_\mu(y') = |D(y', y'')|$.

(h2) tells us that along a path from m to ξ, D_μ is strictly increasing, so $D_\mu(x') = |D(x', x'')| < |D(y', y'')| = D_\mu(y')$. QED.

Note that the Euclidean distance function violates both (h1) and (h2) in all dimensions. The Kullback-Leibler divergence fulfills them if $n = 2$ but violates them if $n > 2$. For $n = 2$, this is easily checked by considering the derivative of the Kullback-Leibler divergence for two dimensions (use D_ε defined in (46) and (47) for the two-dimensional case instead of D_μ for the arbitrary-dimensional case). A counterexample to fulfillment of (h1) and (h2) for $n = 3$ is given in section$^{\text{TBD}}$.

Now that we have shown that (h1) and (h2) is equivalent, it is easy to show that J_P, L_P and Z_P fulfill the horizon requirement while M_P, R_P and G_P violate it. The following table of derivatives will do (note that $\varepsilon = y - x$ is fixed while x varies and that these derivatives have to be considered with the absolute value for the various degree of confirmation functions in mind). Column 1 is the name of the candidate confirmation function; column 2 is the function with x and $\varepsilon = y - x$ as arguments; column 3 is the derivative $\partial D(x, \varepsilon)/\partial x$ for $|D(x, \varepsilon)| = D(x, \varepsilon)$.

$M_P(x, y)$	ε	0
$R_P(x, y)$	$\log \frac{x+\varepsilon}{x}$	$-\frac{\varepsilon}{(x+\varepsilon)x}$
$J_P(x, y)$	$\frac{x+\varepsilon}{1-x-\varepsilon} - \frac{x}{1-x}$	$\frac{1}{(1-x-\varepsilon)^2} - \frac{1}{(1-x)^2}$
$L_P(x, y)$	$\log \frac{(x+\varepsilon)(1-x)}{x(1-x-\varepsilon)}$	(73)
$G_P(x, y)$	$\log \frac{1-x}{1-x-\varepsilon}$	$\frac{h}{(1-x)(1-x-\varepsilon}$
$Z_P(x, y)$	$\frac{\varepsilon}{1-x}$ or $\frac{\varepsilon}{x}$	$\frac{\varepsilon}{(1-x)^2}$ or $-\frac{\varepsilon}{x^2}$

with

$$-\frac{((x+\varepsilon)(1-x) - (1-x-\varepsilon)x)(1-2x-\varepsilon)}{(x+\varepsilon)(1-x-\varepsilon)(1-x)x}. \tag{73}$$

D Appendix: The Hermitian Form Model

Asymmetric MDS is a promising approach to classify asymmetries in terms of their behaviour. This subsection demonstrates that Chino's asymmetric

MDS, both spatial and non-spatial, fails to give us explanations for information theory's violation of TRANSITIVITY OF ASYMMETRY. I am choosing Chino's approach because it is the most general and most promising of all the different asymmetric MDS models (see, for example, Chino and Shiraiwa, 1993, where Chino manages to subsume many of the other approaches into his own).

Multi-dimensional scaling (MDS) visualizes similarity of individuals in datasets. Various techniques are used in information visualization, in particular to display the information contained in a proximity matrix. When the proximity matrix is asymmetrical, we speak of asymmetric MDS. These techniques can be spatial (see for example Chino, 1978), where the proximity relationships are visualized in two-dimensional or higher-dimensional space; or non-spatial (see for example Chino and Shiraiwa, 1993), where the proximity relationships are used to identify data sets with abstract spaces (in Chino's case, finite-dimensional complex Hilbert spaces) and metrics defined on them.

The spatial approach in two dimensions fails right away for information theory because it cannot visualize transitivity violations. The hope for other types of asymmetric MDS is that it would be able to distinguish between well-behaved and ill-behaved asymmetries and either exclude or identify better-behaved candidates than the Kullback-Leibler divergence for measuring the distance between probability distributions. I will use Chino's most sophisticated non-spatial account to show that asymmetric MDS cannot solve this problem. For other asymmetric MDS note that with the Hermitian Form Model Chino seeks to integrate and generalize over all the other accounts.

Assume a finity proximity matrix. I will work with two examples here to avoid the detailed and abstract account provided by Chino. The first example is

$$D = \begin{bmatrix} 0 & 2 & 3 \\ 3 & 0 & 1 \\ -1 & 2 & 0 \end{bmatrix} \tag{74}$$

and allows for easy calculations. The second example corresponds to (61), the example for transitivity violation where

$$\hat{D} = \begin{bmatrix} 0.0000 & 0.0566 & 0.0487 \\ 0.0589 & 0.0000 & 0.0499 \\ 0.0437 & 0.0541 & 0.0000 \end{bmatrix}, \tag{75}$$

and the elements of the matrix $\hat{d}_{jk} = D_{\mathrm{KL}}(P_j, P_k)$. Note that the diagonal elements are all zero, as no updating is necessary to keep the probability distribution constant.

Chino first defines a symmetric matrix S and a skew-symmetric matrix T corresponding to the proximity matrix such that $D = S + T$.

$$S = \frac{1}{2}(D + D') \text{ and } T = \frac{1}{2}(D - D'). \tag{76}$$

Note that D' is the transpose of D, S is a symmetric matrix, and T is a skew-symmetric matrix with $t_{jk} = -t_{kj}$. Next we define the Hermitian matrix

$$H = S + iT, \tag{77}$$

where i is the imaginary unit. H is a Hermitian matrix with $h_{jk} = \overline{h_{kj}}$. Hermitian matrices are the complex generalization of real symmetric matrices. They have special properties (see section 8.9 in Anton and Busby, 2003) which guarantee the existence of a unitary matrix U such that

$$H = U\Lambda U^*, \tag{78}$$

where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ with n the dimension of D and λ_k the k-th eigenvalue of H (theorem 8.9.8 in Anton and Busby, 2003). U is the matrix of eigenvectors with the k-th column being the k-th eigenvector. U^* is the conjugate transpose of U. Given example (74), the numbers look as follows:

$$H = \frac{1}{2}\begin{bmatrix} 0 & 5-i & 2+4i \\ 5+i & 0 & 3-i \\ 2-4i & 3+i & 0 \end{bmatrix} \tag{79}$$

and

$$U = \begin{bmatrix} 0.019 + 0.639i & -0.375 + 0.195i & 0.514 + 0.386i \\ 0.279 - 0.494i & -0.169 - 0.573i & 0.503 + 0.260i \\ -0.519 + 0.000i & 0.681 - 0.000i & 0.516 + 0.000i \end{bmatrix} \tag{80}$$

with $\Lambda = \mathrm{diag}(-3.78, 0.0715, 3.71)$. Λ is calculated using the characteristic polynomial $\lambda^3 - 14\lambda + 1$ of H. Notice that the characteristic polynomial is a depressed cubic (the second coefficient is zero), which facilitates computation and will in the end spell the failure of Chino's program for our purposes.

Given example (75), the numbers are

$$\hat{H} = \frac{1}{2} \begin{bmatrix} 0.0000 + 0.0000i & 0.0578 - 0.0011i & 0.046 + 0.003i \\ 0.0578 + 0.0011i & 0.0000 + 0.0000i & 0.052 - 0.002i \\ 0.0462 - 0.0025i & 0.0520 + 0.0021i & 0.000 + 0.000i \end{bmatrix} \tag{81}$$

and

$$\hat{U} = \begin{bmatrix} 0.351 - 0.467i & -0.543 + 0.170i & -0.578 - 0.006i \\ -0.604 + 0.457i & -0.201 - 0.169i & -0.598 + 0.002i \\ 0.290 - 0.000i & 0.779 + 0.000i & -0.555 + 0.000i \end{bmatrix} \tag{82}$$

with $\Lambda = \mathrm{diag}(-0.060, -0.045, 0.104)$.

Chino now elegantly shows how the decomposition of $H = U\Lambda U^*$ defines a seminorm on a vector space. Let $\phi(\zeta, \tau) = \zeta\Lambda\tau^*$. Then (i) $\phi(\zeta_1 + \zeta_2, \tau) = \phi(\zeta_1, \tau) + \phi(\zeta_2, \tau)$, (ii) $\phi(a\zeta, \tau) = a\phi(\zeta, \tau)$, and (iii) $\phi(\zeta, \tau) = \overline{\phi(\tau, \zeta)}$. These three conditions characterize an inner product on a finite-dimensional complex Hilbert space, but only if a fourth condition is met: positive (or negative) definiteness ($\phi(\zeta, \zeta) \geq 0$) for all ζ). One might hope that positive definiteness identifies the more well-behaved asymmetries by associating with them a finite-dimensional complex Hilbert space with the norm $\|\zeta\| = \sqrt{\phi(\zeta, \zeta)}$ defined on it (Chino himself speculatively mentioned this hope to me in personal communication).

The hope does not come to fruition. Without a non-trivial self-similarity relation, all seminorms defined as above are indefinite, and thus all cats grey in

the night. Not only are well-behaved and ill-behaved asymmetries indistinguishable by the light of this seminorm, even the seminorms for symmetry are indefinite. Not only does this not help our programme, it also puts a serious damper on Chino's, who never mentions the self-similarity requirement (which, given that we are dealing with a proximity matrix, is substantial).

Based on a theorem in linear algebra (see theorem 4.4.12 in Anton and Busby, 2003),

$$\sum_{j=1}^{n} \lambda_j = \mathrm{tr}(A) \tag{83}$$

whenever the λ_j are the eigenvalues of A. The reader can easily verify this theorem by noticing that the roots of the characteristic polynomial add up to the second coefficient (which is the trace of the original matrix). It is well-known that the eigenvalues of a Hermitian matrix are real-valued (theorem 8.9.4 in Anton and Busby, 2003), which is an important component for Chino to define the seminorm $\|\zeta\|$ with the help of ϕ. Unfortunately, using (83), the eigenvalues are not only real, but also add up to the trace of H, which is zero unless there is a non-trivial self-similarity relation.

Tversky entertains such self-similarity relations in psychology (see tbd), and Chino is primarily interested in applications in psychology. When the eigenvalues add up to zero, however, there will be positive and negative eigenvalues (unless the whole proximity matrix is the null-matrix), which renders the seminorm as defined by Chino indefinite. The Kullback-Leibler divergence is trivial with respect to self-similarity: $D_{\mathrm{KL}}(P, P) = 0$ for all P.

References

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

Amari, Shun-ichi, and Hiroshi Nagaoka. *Methods of Information Geometry*. Providence, RI: American Mathematical Society, 2000.

Anton, Howard, and Robert Busby. *Contemporary Linear Algebra*. New York, NY: Wiley, 2003.

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

Atkinson, David. "Confirmation and Justification." *Synthese* 184, 1: (2012) 49–61.

Calin, Ovidiu, and Constantin Udriste. *Geometric Modeling in Probability and Statistics*. Heidelberg, Germany: Springer, 2014.

Carnap, Rudolf. *The Continuum of Inductive Methods*. University of Chicago, 1952.

———. *Preface to the Second Edition of Logical Foundations of Probability*. University of Chicago, 1962.

Caticha, Ariel, and Adom Giffin. "Updating Probabilities." In *MaxEnt 2006, the 26th International Workshop on Bayesian Inference and Maximum Entropy Methods*. 2006.

Chentsov, Nikolai. *Statistical Decision Rules and Optimal Inference*. Providence, R.I: American Mathematical Society, 1982.

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.

Christensen, David. "Measuring Confirmation." *The Journal of Philosophy* 96, 9: (1999) 437–461.

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

Cover, T.M., and J.A. Thomas. *Elements of Information Theory*, volume 6. Hoboken, NJ: Wiley, 2006.

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

Crupi, Vincenzo, and Katya Tentori. "State of the Field: Measuring Information and Confirmation." *Studies in History and Philosophy of Science Part A* 47: (2014) 81–90.

Crupi, Vincenzo, Katya Tentori, and Michel Gonzalez. "On Bayesian Measures of Evidential Support: Theoretical and Empirical Issues." *Philosophy of Science* 74, 2: (2007) 229–252.

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.

Earman, John. *Bayes or Bust?* Cambridge, MA: MIT, 1992.

Festa, R. "Bayesian Confirmation." In *Experience, Reality, and Scientific Explanation: Essays in Honor of Merrilee and Wesley Salmon*, edited by Merrilee H. Salmon, Maria Carla Galavotti, and Alessandro Pagnini, Dordrecht: Kluwer, 1999, 55–88.

Fitelson, Branden. "A Bayesian Account of Independent Evidence with Applications." *Philosophy of Science* 68, 3: (2001) 123–140.

———. "Logical Foundations of Evidential Support." *Philosophy of Science* 73, 5: (2006) 500–512.

Gaifman, Haim. "Subjective Probability, Natural Predicates and Hempel's Ravens." *Erkenntnis* 14, 2: (1979) 105–147.

Gemes, Ken. "Verisimilitude and Content." *Synthese* 154, 2: (2007) 293–306.

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.

Good, Irving. *Good Thinking: The Foundations of Probability and Its Applications*. Minneapolis, MN: University of Minnesota, 1983.

Good, John Irving. *Probability and the Weighing of Evidence*. London, UK: Griffin, 1950.

Greaves, Hilary, and David Wallace. "Justifying Conditionalization: Conditionalization Maximizes Expected Epistemic Utility." *Mind* 115, 459: (2006) 607–632.

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

Hájek, Alan, and James Joyce. "Confirmation." In *Routledge Companion to the Philosophy of Science*, edited by S. Psillos, and M. Curd, New York, NY: Routledge, 2008, 115–129.

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.

Keynes, John Maynard. *A Treatise on Probability*. London, UK: Macmillan, 1921.

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

Kyburg, Henry. "Recent Work in Inductive Logic." In *Recent Work in Philosophy*, edited by Tibor Machan, and Kenneth Lucey, Totowa, NJ: Rowman and Allanheld, 1983, 87–150.

Leitgeb, Hannes, and Richard Pettigrew. "An Objective Justification of Bayesianism I: Measuring Inaccuracy." *Philosophy of Science* 77, 2: (2010a) 201–235.

———. "An Objective Justification of Bayesianism II: The Consequences of Minimizing Inaccuracy." *Philosophy of Science* 77, 2: (2010b) 236–272.

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

Lukits, Stefan. "The Principle of Maximum Entropy and a Problem in Probability Kinematics." *Synthese* 1–23.

———. "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.

Maher, Patrick. *Betting on Theories*. Cambridge University, 1993.

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.

Milne, Peter. "log[P(h/eb)/P(h/b)] Is the One True Measure of Confirmation." *Philosophy of Science* 63, 1: (1996) 21–26.

———. "Information, Confirmation, and Conditionals." *Journal of Applied Logic* 12, 3: (2014) 252–262.

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

Nozick, Robert. *Philosophical Explanations*. Cambridge, MA: Harvard University, 1981.

Oddie, Graham. "The Content, Consequence and Likeness Approaches to Verisimilitude: Compatibility, Trivialization, and Underdetermination." *Synthese* 190, 9: (2013) 1647–1687.

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.

Pettigrew, Richard. "Epistemic Utility and Norms for Credences." *Philosophy Compass* 8, 10: (2013) 897–908.

Popper, Karl. *Conjectures and Refutations*. London, UK: Routledge and Kegan Paul, 1963.

Rosenkrantz, R.D. "Bayesian Confirmation: Paradise Regained." *British Journal for the Philosophy of Science* 45, 2: (1994) 467–476.

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

Salmon, Wesley. "Confirmation and Relevance." In *Induction, Probability, and Confirmation*, edited by Grover Maxwell, and Robert Milford Anderson, Minneapolis, MI: University of Minnesota Press, 1975, 3–36.

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

Shogenji, Tomoji. "The Degree of Epistemic Justification and the Conjunction Fallacy." *Synthese* 184, 1: (2012) 29–48.

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.

Sober, Elliott. "No Model, No Inference: A Bayesian Primer on the Grue Problem." In *Grue!: The New Riddle of Induction*, edited by Douglas Frank Stalker, Chicago: Open Court, 1994, 225–240.

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.

Williams, Paul. "Bayesian Conditionalisation and the Principle of Minimum Information." *The British Journal for the Philosophy of Science* 31, 2: (1980) 131–144.

Zalabardo, José. "An Argument for the Likelihood-Ratio Measure of Confirmation." *Analysis* 69, 4: (2009) 630–635.

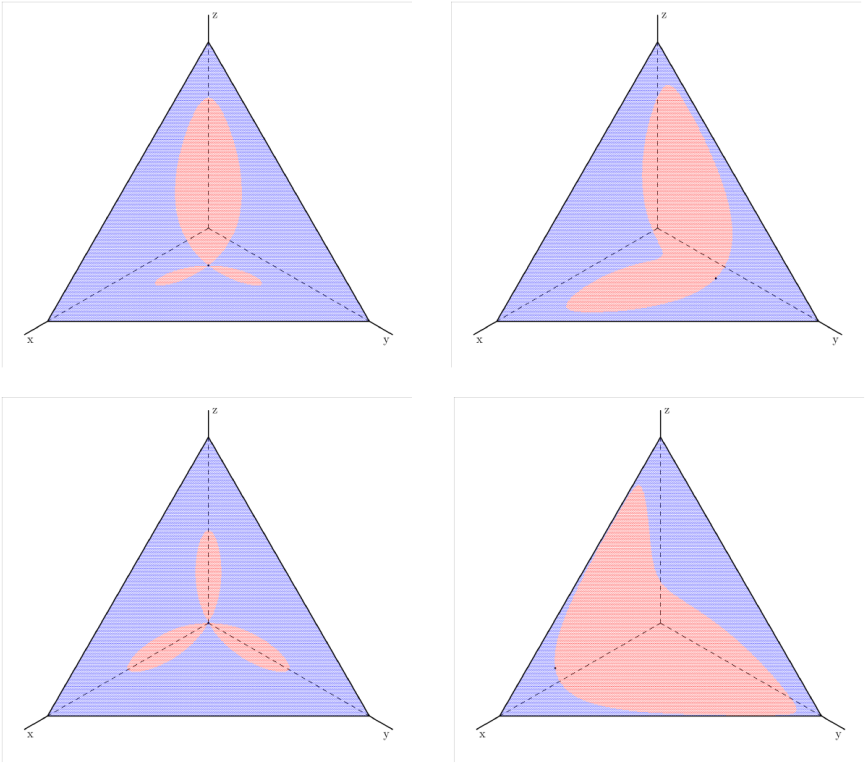

Figure 8: The partition (60) based on different values for P. From top left to bottom right, $P = (0.4, 0.4, 0.2); P = (0.242, 0.604, 0.154); P = (1/3, 1/3, 1/3); P = (0.741, 0.087, 0.172)$. Note that for the geometry of reason, the diagrams are trivial. The challenge for information theory is to explain the non-triviality of these diagrams epistemically without begging the question.