Asymmetry and the Geometry of Reason

Stefan Lukits

1 Introduction

The 'geometry of reason' (a term coined by Richard Pettigrew and Hannes Leitgeb, two of its advocates) 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. This paper demonstrates in detail how it 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 3, 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.

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).

I will show that LP conditioning, which is an alternative to Jeffrey conditioning as a generalization of standard conditioning and which the geometry

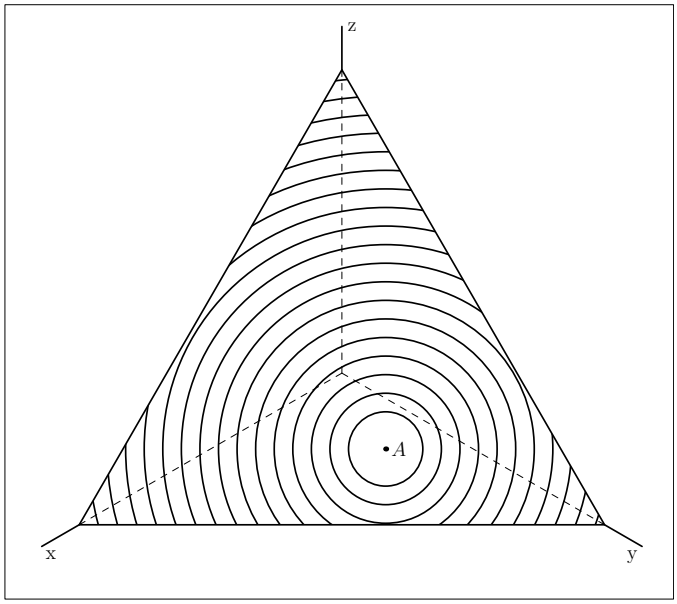

Figure 1: The simplex $\mathbb{S}^2$ in three-dimensional space $\mathbb{R}^3$ with contour lines corresponding to the geometry of reason around point A in equation (1). Points on the same contour line are equidistant from A with respect to the Euclidean metric. Compare the contour lines here to figure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

of reason entails, fails commonsense expectations that are reasonable to have for the kind of updating scenario that LP conditioning addresses. The failure of Jeffrey conditioning, which fulfills these commonsense expectations, to minimize inaccuracy on the basis of the geometry of reason casts, by reductio, doubt on the geometry of reason.

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

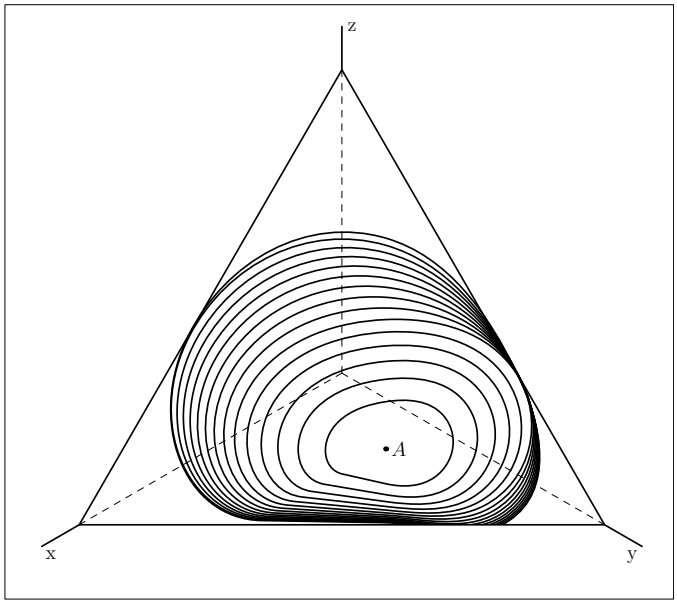

Figure 2: The simplex $\mathbb{S}^2$ with contour lines corresponding to information theory around point A in equation (1). Points on the same contour line are equidistant from A with respect to the Kullback-Leibler divergence. The contrast to figure 1 will become clear in much more detail in the body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

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.

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 in his article "Geometry of Logic and Truth Approximation."

Using an axiomatic approach based on the geometry of reason, Leitgeb and Pettigrew show 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.

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 3.7, 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.

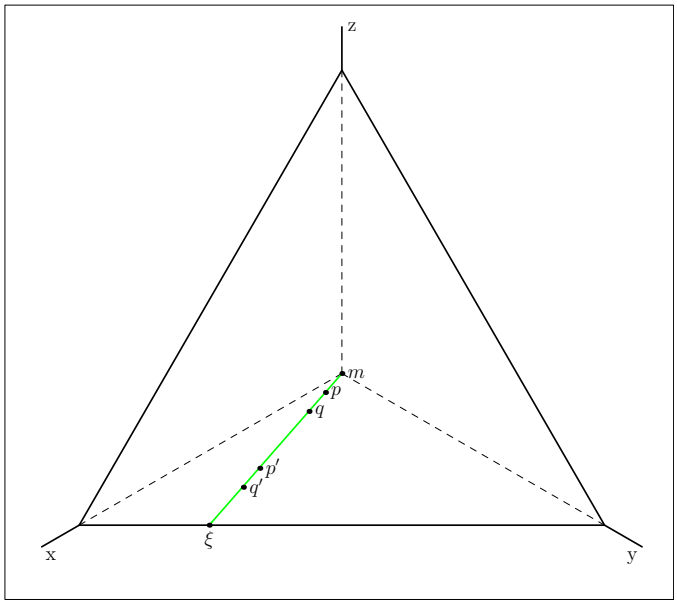

Figure 3: An illustrations of conditions (i)–(iii) for COLLINEAR HORIZON in List Two. p, p' and q, q' must be equidistant and collinear with m and ξ. If q, q' is more peripheral than p, p', then COLLINEAR HORIZON requires that $|d(p, p')| < |d(q, q')|$.

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. 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.

2 Geometry of Reason versus Information Theory

Consider the following three points in three-dimensional space:

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

All three are elements of the simplex $\mathbb{S}^2$: their coordinates add up to 1. Thus they represent probability distributions A, B, C over a partition of the event space into three 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:

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

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

Let the Euclidean distance $\|B-A\|$ be defined as usual by $\sqrt{\sum_{i=1}^{n} (a_i - b_i)^2}$. What is remarkable about the three points in (1) is that

$$\|C - A\| \approx 0.204 < \|B - A\| \approx 0.212 \tag{3}$$

and

$$D_{\mathrm{KL}}(B, A) \approx 0.0589 < D_{\mathrm{KL}}(C, A) \approx 0.069. \tag{4}$$

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

What Leitgeb and Pettigrew call 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.

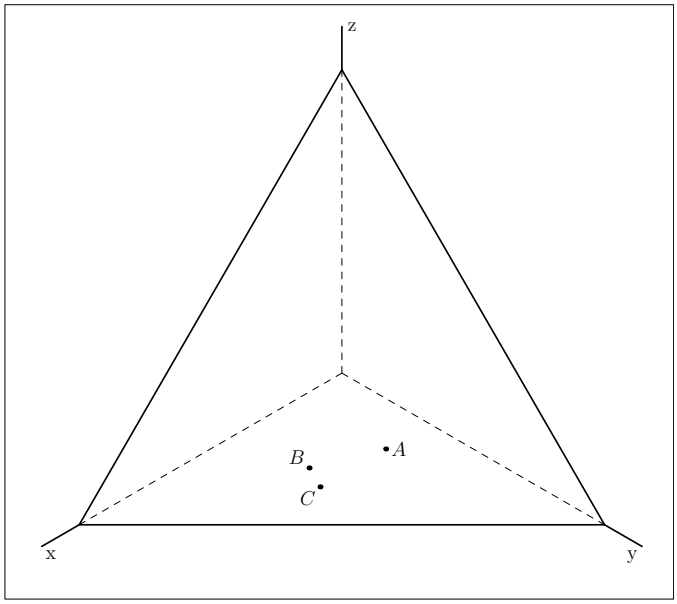

Figure 4: The simplex $\mathbb{S}^2$ in three-dimensional space $\mathbb{R}^3$ with points a, b, c as in equation (1) representing probability distributions A, B, C. Note that geometrically speaking C is closer to A than B is. Using the Kullback-Leibler divergence, however, B is closer to A than C is.

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

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

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

$$\begin{aligned} P(E_1) &= 1/3 \\ P(E_2) &= 1/2 \\ P(E_3) &= 1/6 \end{aligned} \tag{5}$$

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,

$$\begin{aligned} P'_{\mathrm{JC}}(E_i) &= P'(E_i|F^*)P'(F^*) + P'(E_i|F^{**})P'(F^{**}) \\ &= P(E_i|F^*)P'(F^*) + P(E_i|F^{**})P'(F^{**}) \end{aligned} \tag{6}$$

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:

$$\begin{aligned} P(E_1) + x &= P'(F^*) \\ P(E_2) + y + P(E_3) + y &= P'(F^{**}) \end{aligned} \tag{7}$$

and then set

$$\begin{aligned} P'_{\text{LP}}(E_1) &= P(E_1) + x \\ P'_{\text{LP}}(E_2) &= P(E_2) + y \\ P'_{\text{LP}}(E_3) &= P(E_3) + y \end{aligned} \tag{8}$$

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

$$\begin{aligned} P'_{\text{LP}}(E_1) &= 1/2 \\ P'_{\text{LP}}(E_2) &= 5/12 \\ P'_{\text{LP}}(E_3) &= 1/12 \end{aligned} \tag{9}$$

Compare these results to the results of Jeffrey conditioning:

$$\begin{aligned} P'_{\text{JC}}(E_1) &= 1/2 \\ P'_{\text{JC}}(E_2) &= 3/8 \\ P'_{\text{JC}}(E_3) &= 1/8 \end{aligned} \tag{10}$$

Note that (5), (10), and (9) correspond to A, B, C in (1).

3 Expectations for the Geometry of Reason

It remains to provide more detail for the expectations in List One (see page 4) and to show how LP conditioning violates them. These subsections have been abridged to accommodate the word limit for this submission. The full-length paper contains the complete version of these arguments, especially their formal components and examples.

3.1 Continuity

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

3.2 Regularity

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

The violation becomes egregious 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.

3.3 Levinstein

LP conditioning violates LEVINSTEIN because of "the potentially dramatic effect [LP conditioning] can have on the likelihood ratios between different propositions" (Levinstein, 2012, 419). Levinstein proposes a logarithmic inaccuracy measure as a remedy to avoid violation of LEVINSTEIN (vaguely related to the Kullback-Leibler divergence), but his account falls far short of the formal scope, substance, and integrity of information theory. As a special case of applying a Levinstein-type logarithmic inaccuracy measure, information theory does not violate LEVINSTEIN.

3.4 Invariance

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

3.5 Expansibility

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

$$\begin{array}{rcl} P(X_1) & = & 0 \\ P(X_2) & = & 0.3 \\ P(X_3) & = & 0.6 \\ P(X_4) & = & 0.1 \end{array} \tag{11}$$

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 (11) or on the following list:

$$\begin{array}{rcl} P(X_2) & = & 0.3 \\ P(X_3) & = & 0.6 \\ P(X_4) & = & 0.1 \end{array} \tag{12}$$

Whether you update on (11) or (12) 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.

3.6 Horizon

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

> **Example 3: Airplane Gliders.** Compare two scenarios. In the first, an airplane which is considered safe (probability of crashing is $1/10^9$) goes through an inspection where a mechanical problem is found which increases the probability of a crash to 1/100. In the second, military gliders land behind enemy lines, where their risk of perishing is 26%. A slight change in weather pattern increases this risk to 27%. (Schlesinger, 1995, 211)

I claim that an amujus ought to fulfill the requirements of the horizon effect: it ought to be more difficult to update as probabilities become more extreme (or less middling). I have formalized this requirement in List Two (see page 5). It is trivial that the geometry of reason does not fulfill it. Information theory fails as well, which gives the horizon effect its prominent place in both lists. The way information theory fails, however, is quite different. Near the boundary of $\mathbb{S}^{n-1}$, information theory reflects the horizon effect just as our expectation requires. The problem is near the centre, where some equidistant points are more divergent the closer they are to the middle. I will give an example and more explanation in subsection 4.2.

3.7 Confirmation

This is a lengthy subsection and has been cut from the submission in order to accommodate the word limit for this submission. Although it contributes an interesting perspective to the problem, it is not essential to my claims.

4 Expectations for Information Theory

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

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

4.1 Triangularity

The three points A, B, C in (1) violate TRIANGULARITY:

$$D_{\mathrm{KL}}(A, C) > D_{\mathrm{KL}}(B, C) + D_{\mathrm{KL}}(A, B). \tag{13}$$

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:

$$\delta(x, z) > \delta\left(x, y\right) + \delta\left(y, z\right). \tag{14}$$

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

$$T_j = \sum_k \delta\left(y^{jk}, y^{j(k+1)}\right) \tag{15}$$

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 (14), 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 (14), it is straightforward to see that (14) is equivalent to

$$\sum_{i=1}^{n}(z_i - x_i)\log\frac{\vartheta x_i + (1-\vartheta)z_i}{x_i} > 0. \tag{16}$$

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

$$\sum_{i=1}^{n}\left(z_i + \frac{\vartheta}{1-\vartheta}x_i - \frac{\vartheta}{1-\vartheta}x_i - x_i\right)\log\frac{\frac{1}{1-\vartheta}\left(\vartheta x_i + (1-\vartheta)z_i\right)}{\frac{1}{1-\vartheta}x_i} > 0. \tag{17}$$

(17) is clearly equivalent to (16). It is also equivalent to

$$\sum_{i=1}^{n}\left(z_i + \frac{\vartheta}{1-\vartheta}x_i\right)\log\frac{z_i + \frac{\vartheta}{1-\vartheta}x_i}{\frac{1}{1-\vartheta}x_i} + \sum_{i=1}^{n}\frac{1}{1-\vartheta}x_i\log\frac{\frac{1}{1-\vartheta}x_i}{z_i + \frac{\vartheta}{1-\vartheta}x_i} > 0, \tag{18}$$

which is true by Gibbs' inequality.

4.2 Collinear Horizon

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

$$p = \left(\frac{1}{5}, \frac{2}{5}, \frac{2}{5}\right) \qquad p' = q = \left(\frac{1}{4}, \frac{3}{8}, \frac{3}{8}\right) \qquad q' = \left(\frac{3}{10}, \frac{7}{20}, \frac{7}{20}\right) \tag{19}$$

The conditions of COLLINEAR HORIZON in List Two (see page 5) 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,

$$D_{\mathrm{KL}}(B, A) = 7.3820 \cdot 10^{-3} > 6.4015 \cdot 10^{-3} = D_{\mathrm{KL}}(C, B). \tag{20}$$

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 one could attenuate the independent appeal of this violation against information theory is by making it parasitic on the asymmetry of information theory, but the details are beyond the scope of this 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.

4.3 Transitivity of Asymmetry

Asymmetry presents a problem for the geometry of reason as well as for information theory. For the geometry of reason, the problem is akin to CON-

TINUITY. 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 4: 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

$$\Delta_P(P') = D_{\mathrm{KL}}(P', P) - D_{\mathrm{KL}}(P, P'), \tag{21}$$

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

$$\Delta^{-1}(\mathbb{R}_{>0}) \qquad \Delta^{-1}(\mathbb{R}_{<0}) \qquad \Delta^{-1}(\{0\}) \tag{22}$$

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

In higher-dimensional cases, however, the tripartite partition (22) 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.

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)