Asymmetry and the Geometry of Reason

Anonymous

1 Introduction

The geometry of reason refers to a view of epistemic utility in which the underlying topology for credence functions (which may be subjective probability distributions) on a finite number of events is a metric space. 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).

My claim is that given an epistemic utility approach and some intuitive axioms, the geometry of reason leads itself ad absurdum; and that there is a viable alternative to the geometry of reason which avoids the problematic implications: information theory. 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).

2 Epistemic Utility and the Geometry of Reason

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, the geometry of reason is adopted without much reflection. Terms such as 'midpoint' between two distributions and $\lambda b' + (1-\lambda)b''$ for distributions 'between' two distributions b' and b'' are used freely.

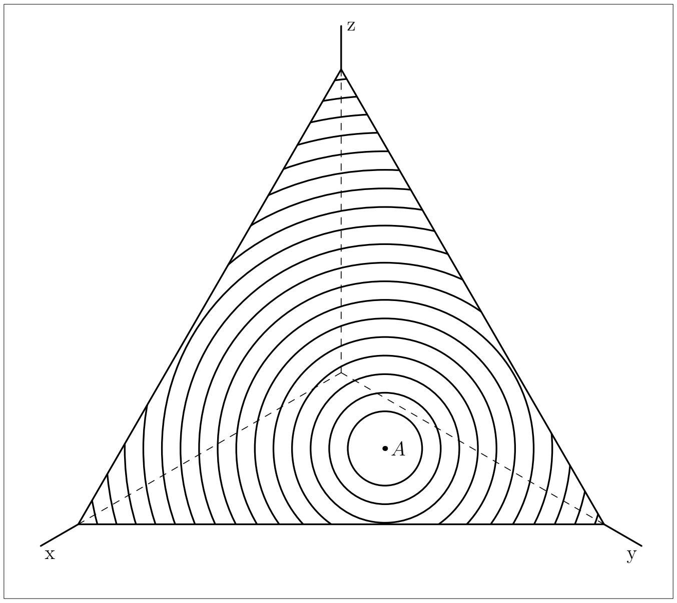

Figure 1: The simplex $\mathbb{S}^3$ in three-dimensional space $\mathbb{R}^3$ with contour lines corresponding to the geometry of reason around point A in equation (1). Points on the same contour line are equidistant from A with respect to the Euclidean metric. Compare the contour lines here to figure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

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 Shun-ichi, 1985).

Joyce advocates for axioms such as Weak Convexity and Symmetry 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 subsection 4.7, I will show that Weak Convexity holds, and Symmetry does not hold, in 'information geometry.' Information theory in this context is the topology generated by the Kullback-Leibler divergence. The term information geometry is due to Imre Csiszár, who considers the Kullback-Leibler divergence an asymmetric analogue of squared Euclidean

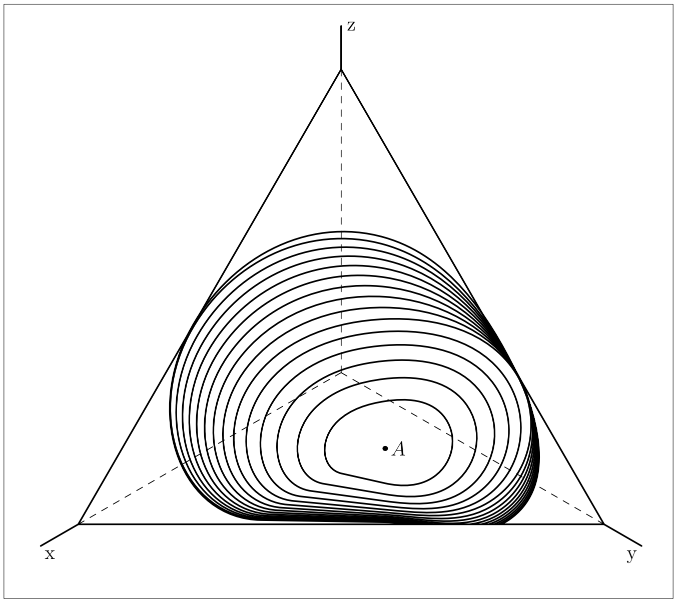

Figure 2: The simplex $\mathbb{S}^3$ with contour lines corresponding to information theory around point A in equation (1). Points on the same contour line are equidistant from A with respect to the Kullback-Leibler divergence. The contrast to figure (1) will become clear in much more detail in the body 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. The main argument of 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.

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).

Leitgeb and Pettigrew's work is continuous with Joyce's work, although their axioms tend to be stronger including expected inaccuracies. They show that uniform distribution (a version of the principle of indifference) 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, 2010, 250f); and that Jeffrey conditioning does not fulfill Joyce's Norm of Gradational Accuracy (see Joyce, 1998, 579). 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, we note that LP conditioning violates all of the following seven plausible expectations for an amujus, an 'alternative method of updating for Jeffrey-type updating scenarios.'

- 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).
- INVARIANCE An amujus ought to be partition invariant.
- 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.
- CONFIRMATION An amujus ought to align with intuitions we have about degrees of confirmation.
- ASYMMETRY An amujus ought to reflect epistemic asymmetries. Updating from one probability distribution to another may need to be reflected in a different proximity relation than going the opposite way.

We are faced with the choice of rejecting the geometry of reason or accepting these unpleasant consequences. Fortunately, there is a live alternative to the geometry of reason: information theory. Information theory has its own axiomatic approach to justifying probabilism and standard conditioning (see Shore and Johnson, 1980). Furthermore, information theory provides a justification for Jeffrey conditioning and generalizes it (see Lukits, 2015). Information theory is not a geometry of reason because it measures divergences, not distances, between distributions of partial belief. The divergence of b'' from b' may not be equal to the divergence of b' from b''. Updating methods based on information theory (standard conditioning, Jeffrey conditioning, the principle of maximum entropy) fulfill the seven expectations.

The next section provides 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 seven expectations outlined above. The conclusion is that a rational agent uses information theory, not the geometry of reason.

3 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 three-dimensional simplex $\mathbb{S}^3$: their coordinates add up to 1. Thus they represent probability distributions over a partition of the event space into three events. Now call $D_{\text{KL}}(A,B)$ the Kullback-Leibler divergence of A from B defined as follows, where a_i are the Cartesian coordinates of A:

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

The Euclidean distance $\|A-B\|$ is defined as

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

What is remarkable about the three points in (1) is that

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

and

$$D_{\text{KL}}(A,B) \approx 0.057 < D_{\text{KL}}(A,C) \approx 0.072. \tag{5}$$

The Kullback-Leibler divergence and Euclidean distance give different recommendations with respect to proximity (for illustration see figure 3). If A corresponds to my prior and my evidence is such that I must change the first coordinate to $1/2$ and nothing stronger, then information theory via the Kullback-Leibler divergence recommends the posterior corresponding to B, whereas the geometry of reason recommends the posterior corresponding to C.

Here is a brief outline how Leitgeb and Pettigrew arrive at posterior probability distributions in Jeffrey-type updating scenarios, using their invariance criterion with respect to global and local inaccuracy. 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

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

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

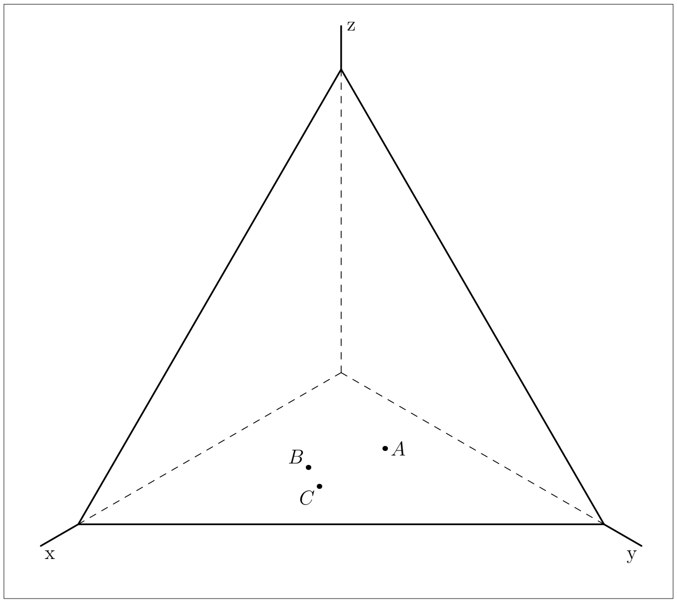

Figure 3: The simplex $\mathbb{S}^3$ in three-dimensional space $\mathbb{R}^3$ with points A, B, C as in equation (1). Note that geometrically speaking C is closer to A than B is. Using the Kullback-Leibler divergence, however, B is closer to A than C is. The reason is asymmetry in information theory, which accords with our intuitions about epistemic space.

Jeffrey conditioning works on the intuition that 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):

$$\begin{aligned} P'_{\text{JC}}(E_i) &= P'(E_i|F^*)P'(F^*) + P'(E_i|F^{**})P'(F^{**}) \\ &= P(E_i|F^*)P'(F^*) + P(E_i|F^{**})P'(F^{**}) \end{aligned} \tag{7}$$

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:

$$\begin{array}{rcl} P(E_1) + x & = & P'(F^*) \\ P(E_2) + y + P(E_3) + y & = & P'(F^{**}) \end{array} \tag{8}$$

and then set

$$\begin{array}{rcl} P'_{\text{LP}}(E_1) & = & P(E_1) + x \\ P'_{\text{LP}}(E_2) & = & P(E_2) + y \\ P'_{\text{LP}}(E_3) & = & P(E_3) + y \end{array} \tag{9}$$

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

$$\begin{array}{rcl} P'_{\text{LP}}(E_1) & = & 1/2 \\ P'_{\text{LP}}(E_2) & = & 5/12 \\ P'_{\text{LP}}(E_3) & = & 1/12 \end{array} \tag{10}$$

Compare these results to the results of Jeffrey conditioning:

$$\begin{array}{rcl} P'_{\text{JC}}(E_1) & = & 1/2 \\ P'_{\text{JC}}(E_2) & = & 3/8 \\ P'_{\text{JC}}(E_3) & = & 1/8 \end{array} \tag{11}$$

Note that (6), (11), and (10) correspond to A, B, C in (1).

4 Seven Expectations

It remains to provide more detail for the seven expectations 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.

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.

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 especially egregious if we are already somewhat 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.

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.). 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.

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.

4.5 Horizon

> **Example 3: 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. Consequently, the step from 89% to 90% is much greater than the step from 67% to 68%.

The emphasis in this argument is on distance, not confirmation, but the next subsection can be considered to be a special case of HORIZON. LP conditioning violates HORIZON because it ignores the epistemic intuition that proximity relations near extreme probabilities are different than away from them (more central rather than peripheral). It should be noted that there are non-Euclidean metrics that obey both HORIZON and CONFIRMATION.

4.6 Confirmation

From an epistemic perspective, updating towards extreme probabilities should become increasingly difficult. Once a hypothesis is already considered to be highly likely or highly unlikely, confirmation or disconfirmation is much harder to come by than in the case of near-equiprobability between alternative hypotheses. The geometry of reason ignores this analogy from confirmation theory; information theory reflects it.

David Christensen's account of degree of confirmation, for example, shows how S-support given by E is stable over Jeffrey conditioning on $\{E, \neg E\}$, which is not the case for LP-conditioning (see Christensen, 1999, 451). LP conditioning violates CONFIRMATION.

4.7 Asymmetry

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

Even the scrupulous about partial beliefs (such as Isaac Levi) concede that 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.

The asymmetry is obvious when extreme probabilities are in play.

Example 4: Extreme Asymmetry. Consider two cases where for case 1 the prior probabilities are $P(Y_1) = 0.4, P(Y_2) = 0.3, P(Y_3) = 0.3$ and the posterior probabilities are $P'(Y_1) = 0, P'(Y_2) = 0.5, P'(Y_3) = 0.5$; for case 2 the prior probabilities are $Q(Y_1) = 0, Q(Y_2) = 0.5, Q(Y_3) = 0.5$ and the posterior probabilities are $Q'(Y_1) = 0.4, Q'(Y_2) = 0.3, Q'(Y_3) = 0.3$;

Case 1 is a straightforward application of standard conditioning. Case 2 is much 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 may be exposed to violating REGULARITY.

Consider now the problem for information theory. 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.

Let me put this differently to emphasize the gravity of the situation for information theory. 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 $\mathbb{S}^3$ represents all of these probability distributions. Every point P in $\mathbb{S}^3$ representing a probability distribution induces a partition on $\mathbb{S}^3$ 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

$$\Delta_P(P') = D_{\mathrm{KL}}(P', P) - D_{\mathrm{KL}}(P, P'), \tag{12}$$

then, holding P fixed, $\mathbb{S}^3$ is partitioned into three regions, $\Delta^{-1}(\mathbb{R}_{>0})$ (red in figure 4), $\Delta^{-1}(\mathbb{R}_{<0})$ (blue in figure 4), and $\Delta^{-1}(\{0\})$ (in figure 4, this would be the line between the red and the blue). 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