Asymmetry and the Geometry of Reason

for blind review

1 Introduction

There are various approaches to justify the ampliative reasoning that goes into choosing a strict subset of the logically and probabilistic coherent partial belief updates to be considered rational. Often, these approaches work hand in hand with each other and are not mutually exclusive. James Joyce, in his paper "A Nonpragmatic Vindication of Probabilism," uses an approach based on epistemic utility rather than pragmatic utility, calibration, or logic of partial belief change (see Joyce, 1998). For Joyce, norms of gradational accuracy characterize the epistemic utility approach to partial beliefs, analogous to norms of truth for full beliefs. The rational agent's reasoning renders inadmissible all of those partial beliefs that are demonstrably inferior in epistemic terms to those partial beliefs left in the updated credal state. Demonstrable inferiority in Joyce is usually based on supervaluationist semantics, if for example one credence is more truth-defective than another credence in all possible worlds.

A set of partial beliefs can be genuinely defective in epistemic terms compared to another set of partial beliefs if its truth estimates are necessarily worse than the truth estimates of its rival. Note that with an estimate, as opposed to a guess, I am not simply right or wrong about the estimated quantity—I am instead more or less accurate. Joyce uses his norm of gradational accuracy together with six axioms (structure, extensionality, dominance, normality, weak convexity, symmetry) to give an epistemic justification of probabilism: the requirement for a rational agent to keep her partial beliefs in keeping with the axioms of probability theory.

It is a natural question to ask whether the same line of reasoning can give us an epistemic justification of standard conditioning and Jeffrey conditioning. In a series of articles that will be pivotal for the rest of this paper, Hannes Leitgeb and Richard Pettigrew show that using Joyce's approach, accuracy can be bifurcated into local and global accuracy. For this bifurcation to give consistent results, the Brier score must be used for Joyce's norm of gradational accuracy. The Brier score vindicates standard conditioning and rules out Jeffrey conditioning. A new type of conditioning, which I shall call LP conditioning, takes the place of Jeffrey conditioning (for details see Leitgeb and Pettigrew, 2010a).

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

If the geometry of reason gives us bad results, we need to look for alternatives. There is a way to relate probability distributions to each other that stands in contrast to the geometry of reason. Its concept of difference is based on information theory, originally conceived to model the extra coding required when the sender gets the probability distribution of the alphabet not quite right. There is a substantial formal theory of information, serendipitously cast in terms of probabilities; there is a similarly substantial formal theory of probabilities and Bayesian epistemology, also cast in terms of probabilities. The formal theories are different: one uses Shannon's entropy, a logarithmic measure of difference; the other uses primarily ratios. It turns out that the superficial differences are rooted in deep commonalities. Information theory can serve as justification for Bayesian norms (probabilism, standard conditioning, and Jeffrey conditioning, for the latter two see Lukits, 2015, 1697f).

Just as there is something special about Euclidean geometry and its attendant Brier scoring rule, there is something special about information theory and its attendant Kullback-Leibler divergence. My results in this paper are largely negative: there are insurmountable problems for the geometry of reason; there are serious problems for information theory. The problems for information theory, however, are such that they deserve attention and possibly have solutions.

The full-length paper contains formal proofs and detailed descriptions with examples of the violations listed in the body of the paper. They have been cut for this submission to fulfill the word count requirement.

1.1 Expectations for Jeffrey-Type Updating Scenarios

For the remainder of this paper I will assume probabilism and an isomorphism between probability distributions P on an outcome space Ω with $|\Omega| = n$ and points $p \in \mathbb{S}^{n-1} \subset \mathbb{R}^n$ having coordinates $p_i = P(\omega_i), i = 1, \ldots, n$ and $\omega_i \in \Omega$. Since the isomorphism is to a metric space, there is a concept of difference between credence functions which can be used to formulate axioms relating credences to epistemic utility and to justify or to criticize contentious positions such as Bayesian conditionalization, the principle of indifference, other forms of conditioning, or probabilism itself (see Joyce, 1998; Leitgeb and Pettigrew, 2010b; and Greaves and Wallace, 2006).

For information theory, as opposed to the geometry of reason, the underlying topology for credence functions is not a metric space (see figures 1 and 2 for illustration). The term information geometry is due to Imre Csiszár, who considers the Kullback-Leibler divergence a non-commutative (asymmetric) analogue of squared Euclidean distance and derives several results that are intuitive information geometric counterparts of standard results in Euclidean geometry (see chapter 3 of Csiszár and Shields, 2004).

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

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

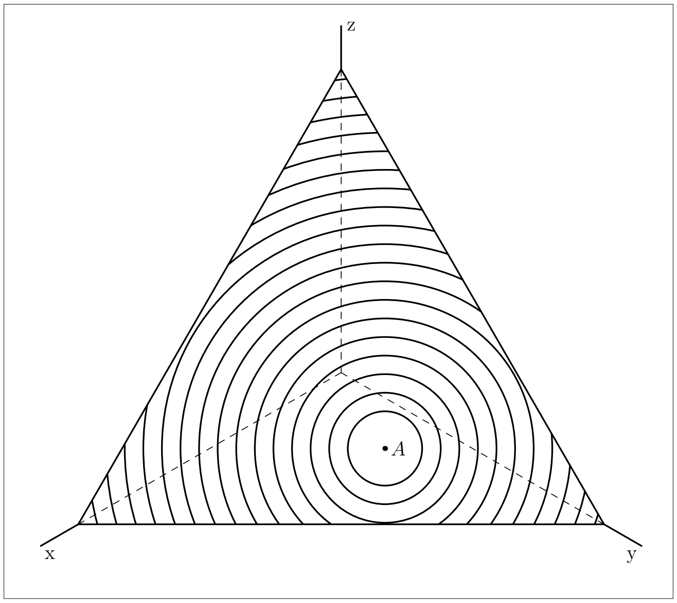

Figure 1: The simplex $\mathbb{S}^2$ in three-dimensional space $\mathbb{R}^3$ with contour lines corresponding to the geometry of reason around point A in equation (1). Points on the same contour line are equidistant from A with respect to the Euclidean metric. Compare the contour lines here to figure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

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

- CONTINUITY An amujus ought to be continuous with standard conditioning as a limiting case.
- REGULARITY An amujus ought not to assign a posterior probability of 0 to an event which has a positive prior probability and about which the intervening evidence says nothing except that a strictly weaker event has a positive posterior probability.
- LEVINSTEIN An amujus ought not to give "extremely unattractive" results in a Levinstein scenario (see Levinstein, 2012, which not only

articulates this failed expectation for LP conditioning, but also the previous two).

- INVARIANCE An amujus ought to be partition invariant.
- EXPANSIBILITY An amujus ought to be insensitive to an expansion of the event space by zero-probability events.
- HORIZON An amujus ought to exhibit the horizon effect which makes probability distributions which are nearer to extreme probability distributions appear to be closer to each other than they really are.

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

- TRIANGULARITY The concept obeys the triangle inequality. If there is an intermediate probability distribution, it will not make the difference smaller: $d(p,r) \leq d(p,q) + d(q,r)$. Buying a pair of shoes is not going to be more expensive than buying the two shoes individually.
- COLLINEAR HORIZON This expecation is just a more technical restatement of the HORIZON expectation in the previous list. If p, p', q, q' are collinear with the centre of the simplex m (whose coordinates are $m_i = 1/n$ for all i) and an arbitrary but fixed boundary point $\xi \in \partial\mathbb{S}^{n-1}$ and p, p', q, q' are all between m and ξ with $\|p'-p\| = \|q'-q\|$ where p is strictly closest to m, then $|d(p,p')| < |d(q,q')|$. For an illustration of this expectation see figure 3.
- TRANSITIVITY OF ASYMMETRY An ordered pair (p,q) of simplex points associated with probability distributions is asymmetrically negative, positive, or balanced, so either $d(p,q)-d(q,p) < 0$ or $d(p,q)-d(q,p) > 0$ or $d(p,q)-d(q,p) = 0$. If (p,q) and (q,r) are asymmetrically positive, (p,r) ought not to be asymmetrically negative. Think of a bicycle route map with different locations at varying altitudes. If it takes 20 minutes

> to get from A to B but only 15 minutes to get from B to A then (A, B) is asymmetrically positive. If (A, B) and (B, C) are asymmetrically positive, then (A, C) ought not to be asymmetrically negative.

While the Kullback-Leibler divergence of information theory fulfills all the expectations of **List A**, save HORIZON, it fails all the expectations in **List B**. Conversely, the Euclidean distance of the geometry of reason fulfills all the expectations of **List B**, save COLLINEAR HORIZON, and fails all the expectations in **List A**.

2 Geometry of Reason versus Information Theory

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

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

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

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

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

The Euclidean distance is defined as follows:

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

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

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

and

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

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

2.1 LP conditioning and Jeffrey Conditioning

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

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

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

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

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

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. Hence,

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

Information theory supports Jeffrey conditioning. Leitgeb and Pettigrew show that Jeffrey conditioning does not in general pick out the minimally inaccurate posterior probability distribution. Leitgeb and Pettigrew introduce an alternative to Jeffrey conditioning, which I call LP conditioning (see Leitgeb and Pettigrew, 2010b, 254. The results for example 2 are:

$$\begin{array}{lcl} P'_{\text{LP}}(E_1) & = & 1/2 \\ P'_{\text{LP}}(E_2) & = & 5/12 \\ P'_{\text{LP}}(E_3) & = & 1/12 \end{array} \tag{8}$$

Compare these results to the results of Jeffrey conditioning:

$$\begin{array}{lcl} P'_{\text{JC}}(E_1) & = & 1/2 \\ P'_{\text{JC}}(E_2) & = & 3/8 \\ P'_{\text{JC}}(E_3) & = & 1/8 \end{array} \tag{9}$$

Note that (6), (9), and (8) correspond to A, B, C in (1).

3 Expectations for the Geometry of Reason

This section provides more detail for the expectations in **List A** and shows how LP conditioning violates them (full-length paper).

4 Expectations for Information Theory

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

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

4.1 Triangularity

The three points A, B, C in (1) violate TRIANGULARITY. Information theory, however, does not only violate TRIANGULARITY. It violates it in a particularly egregious way. Consider any distinct two points x and z on $\mathbb{S}^{n-1}$ with coordinates x_i and z_i $(1 \leq i \leq n)$. For simplicity, let us write $\delta(x, z) = D_{\text{KL}}(z, x)$. Then, for any $\vartheta \in (0, 1)$ and an intermediate point y with coordinates $y_i = \vartheta x_i + (1 - \vartheta) z_i$, the following inequality holds true:

$$\delta(x, z) > \delta\left(x, y\right) + \delta\left(y, z\right). \tag{10}$$

Proof and more detail about the connection of TRIANGULARITY to Bregman divergences are in the full-length paper.

4.2 Collinear Horizon

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

$$p=\left(\frac{1}{5},\frac{2}{5},\frac{2}{5}\right)\qquad p'=q=\left(\frac{1}{4},\frac{3}{8},\frac{3}{8}\right)\qquad q'=\left(\frac{3}{10},\frac{7}{20},\frac{7}{20}\right)\tag{11}$$

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

$$D_{\mathrm{KL}}(B,A)=7.3820\cdot 10^{-3}>6.4015\cdot 10^{-3}=D_{\mathrm{KL}}(C,B).\tag{12}$$

Proof and more detail about the connection of COLLINEAR HORIZON to asymmetry are in the full-length paper.

4.3 Transitivity of Asymmetry

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 geometry of reason may want to solve this problem by signing on to a version of regularity, but then it violates REGULARITY.

Now turn to 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.

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

In other words, if

$$\Delta_P(P') = D_{\text{KL}}(P', P) - D_{\text{KL}}(P, P'), \tag{13}$$

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

$$\Delta^{-1}(\mathbb{R}_{>0}) \qquad \Delta^{-1}(\mathbb{R}_{<0}) \qquad \Delta^{-1}(\{0\}) \tag{14}$$

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 (14) 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 5 for illustration).

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

$$P_1 = \left(\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\right) \qquad P_2 = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right) \qquad P_3 = \left(\frac{2}{5}, \frac{2}{5}, \frac{1}{5}\right) \tag{15}$$

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. This ‘ill behaviour’ of information theory begs for explanation, or at least classification (it would help, for example, to know that all reasonable non-commutative difference measures used for updating are ill-behaved). For a future research project, it would be interesting either to see information theory debunked in favour of an alternative geometry (this paper has demonstrated that this alternative will not be the geometry of reason); or to see uniqueness results for the Kullback-Leibler divergence to show that despite its ill behaviour the Kullback-Leibler is the right asymmetric distance measure on which to base inference and updating.