Asymmetry and the Geometry of Reason

for blind review

Abstract

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. 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 a non-commutative divergence instead of a symmetric distance measure. I show that information theory, despite initial promise, also fails to accommodate basic epistemic intuitions.

1 Introduction

There are various ways in which epistemic norms for partial beliefs are justified. Three important standards of justification in the literature are pragmatics, accuracy, and evi-

dence. A partial belief, as opposed to a full belief, expresses uncertainty whether or not a proposition is true. In formal epistemology, this uncertainty is captured in mathematical models. Examples for epistemic norms are probabilism (the best quantitative model of partial beliefs is such that partial beliefs correspond to probabilities), Bayesian conditionalization (a rational agent updates beliefs using conditional probabilities when possible), the principle of indifference (if there is no further information, then mutually disjoint and jointly exhaustive events distinguishable by name only are equiprobable), and additional updating methods beyond standard conditioning (Jeffrey conditioning, affine constraints). Whether the justification for (or rejection of) these norms cites pragmatic, alethic, or evidential reasons, one important ingredient of the formal model is scoring rules.

This paper investigates scoring rules for partial beliefs that exclusively reward and penalize on epistemic grounds. This aligns my paper roughly with the school that privileges alethic justification for epistemic norms, but it should be noted that the debate over scoring rules also has important implications for pragmatists and evidentialists, who hold that epistemic norms are rooted in decision theory and the appropriate relationship between beliefs and the evidence on which they are based, respectively.

Example 1.1 (Trichotomy)**.** A game between the home team and the away team ends in a win (for the home team), a loss, or a tie.

I am restricting myself to finite algebras of propositions, as in example 1.1, where exactly one of three random outcomes takes place. Let these outcomes (or possible worlds) be ξ_1, ξ_2, ξ_3. Let the agent's "report" over these three outcomes be $c = (c_1, c_2, c_3)^\intercal$. The report may represent a partial belief or an attempt at prediction. The transpose $^\intercal$ symbol merely turns the list of numbers into a vector. Another restriction for this paper is that

the c_i are non-negative real numbers so that the vector c is located in the non-negative orthant $\mathcal{D}_0$ of an $n = 3$ dimensional vector space. The agent is penalized for reporting c according to a loss function which she wants to minimize; the agent does not care for anything outside of what the loss function is able to capture. Her penalty is

$$S(\xi_i, c) \tag{1}$$

once it is established that ξ_i is the realized outcome. $S(c)$ is the vector

$$S(c) = (S(\xi_1, c), \ldots, S(\xi_n, c))^\intercal . \tag{2}$$

To discourage any dishonesty on the agent's part, a common requirement is that the scoring rule be proper. The strict propriety of a scoring rule ensures that an agent reports the same distribution according to which she thinks the random process selects the outcome. Let $\langle ., . \rangle$ be the inner product of two vectors (or the matrix product if dual spaces are used),

$$\langle c, \hat{c} \rangle = \sum_{i=1}^{n} c_i \hat{c}_i. \tag{3}$$

Definition 1.1. A scoring rule S is strictly proper if and only if

$$\langle c, S(c) \rangle < \langle c, S(\hat{c}) \rangle \text{ for all } \hat{c} \in \mathcal{D}_0 \setminus \{c\}. \tag{4}$$

A scoring rule S is proper if and only if (4) is true with a $\leq$ symbol rather than $<$. Strict

propriety guarantees that an agent is motivated to report the distribution that they deem to be the one according to which the random outcomes are generated. Strict propriety significantly narrows down the set of acceptable scoring rules. John McCarthy shows that strict propriety requires the existence of a convex entropy function, for which the scoring rule is a type of derivative (see McCarthy, 1956). Scoring rules which are generated by convex functions are called Bregman divergences (see Bregman, 1967).

The problem I am addressing in this paper is whether there are further restrictions on rationally acceptable scoring rules. McCarthy's theorem leaves open the possibility for symmetric and asymmetric scoring rules. A symmetric scoring rule is associated with a divergence function which assigns as much divergence from a credence c to another credence $\bar{c}$ as vice versa—the divergence function is then more narrowly called a distance function. Reinhard Selten and Richard Pettigrew have recently defended symmetric scoring rules (for example in Selten, 1998; Pettigrew, 2016, 80).

The claim at the heart of my paper is that a defence of symmetry reveals a misapprehension about partial beliefs and their relationships to each other. The misapprehension is that there is a geometry of partial beliefs which is more readily accessible to intuition. It is tempting to view a credence $c = (c_1, c_2, c_3)^\intercal$, for example, as a vector in 3-dimensional space and then evaluate its distance to other credences in terms of its metric distance to them. I will call this view, following Hannes Leitgeb and Richard Pettigrew (see Leitgeb and Pettigrew, 2010a, 210), the geometry of reason.

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 accustomed 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 a Euclidean structure. On the contrary, this would appear to be an improbable coincidence. (Mormann, 2005, 453.)

For Pettigrew, the geometry of reason stands out as an appealing account because its associated scoring rule, the Brier score, is (up to linear transformations) unique once symmetry is required. Pettigrew even has an argument why this uniqueness has appealing features in its own right. I will provide a more detailed summary of the various requirements that one might have with respect to scoring rules. An alternative to the geometry of reason emerges from this analysis: information theory.

Information theory, just like the geometry of reason, has an associated scoring rule: the Log score. The two different scoring rules, Brier score and Log score, agree on recommending probabilism, but they license different updating methods in dynamic partial belief theory. They agree on Bayesian updating in the standard case (using conditional probabilities), but they disagree on updating in a Jeffrey-type updating scenario. Information theory is the better Bayesian, because Bayesian standard conditioning is continuously generalized in information theory to more general updating situations. The more general updating

situations include affine constraints (as in the notorious Judy Benjamin example), of which Jeffrey-type updating scenarios are a special case, and even more complicated constraints where the evidence only allows for a convex region of credence functions. The generalization for the geometry of reason is discontinuous at best, implausible at worst. I will present this view in full detail in the main body of the paper.

The Log score is asymmetric, but there are other asymmetric strictly proper scoring rules. According to Pettigrew's independent argument why it is a good idea to have a unique scoring rule this would count against the Log score and against information theory. The Log score, however, is unique in fulfilling a locality requirement that arguably commands as much plausibility as symmetry. Yet the tenor of my paper is that Pettigrew's independent argument for uniqueness is suspect (in defence of Pettigrew, many of the claims in his book *Accuracy and the Laws of Credence* do not depend on it) and that neither the Brier score's symmetry nor the Log score's locality is sufficient to make them uniquely superior to other scoring rules.

I believe that there are serious problems with the geometry of reason, to the point where I would reject it as a plausible formal account of partial beliefs. As I will show, however, there are serious problems with information theory and how it accommodates epistemic intuitions as well. These are not insurmountable. My hope is that a further detachment from geometry can give us a better understanding of why information theory has the odd features that I will highlight in the paper.

2 Features of Scoring Rules

2.1 List of Features and Preliminaries

Consider the following list of features for a scoring rule SR.

propriety The SR encourages an agent to report the distribution which is her best guess for what generates the random event.

geometry The divergence function associated with the SR is a metric. Consequently, credence functions can be 'visualized' with a distance defined between them.

information The entropy function associated with the SR fulfills Claude Shannon's axioms for an entropy function.

symmetry The divergence function associated with the SR is symmetric.

locality How a distribution scores when an event takes place depends only on the credence assigned by the distribution to this event.

horizon The divergence function associated with the SR has a tendency to measure distributions near the centre as being closer together than distributions near the extremes, all else being equal.

sensitivity The divergence function associated with the SR is sensitive in the sense that distributions strictly between two distributions are closer to the end points than the end points are to each other.

conditioning The SR licenses standard conditioning.

reductio-resistance The SR is not led ad absurdum by licensing implausible updating methods.

univocal dominance The SR is uniquely superior to all other SRs in order to address the Bronfman objection.

Here is a brief summary with evaluative annotation (plausibility, for example, means that in conclusion to my arguments in this paper I find the requirement plausible). If my evaluation is endorsed, only the Log score qualifies as a rationally acceptable scoring rule and information theory (as opposed to the geometry of reason) is vindicated; however, the violation of HORIZON must be explained, which is beyond the purview of this paper.

requirement	**Brier score**	**Log score**	**other scores**	**annotation**
propriety	yes	yes	yes	commonly accepted
geometry	yes	no	no	implausible
information	no	yes	no	plausible
symmetry	yes	no	no	implausible
locality	no	yes	no	weakly plausible
horizon	no	long story	perhaps	plausible
sensitivity	yes	no	perhaps	implausible
conditioning	yes	yes	perhaps	plausible
reductio-resistance	no	yes	perhaps	plausible
univocal dominance	yes	yes	perhaps	implausible

Let $\mathcal{A}$ be an algebra with cardinality $k = 2^m$ over the events in the finite sample space Ω, whose cardinality is m. A credence function over $\mathcal{A}$ is a vector in the positive orthant of

$\mathbb{R}^k$. An orthant generalizes a quadrant in $\mathbb{R}^2$ to $\mathbb{R}^k$. I will write $\mathcal{D}_0$ if the orthant includes vectors which have elements that equal zero; $\mathcal{D}$ if all elements of the vector are greater than zero. The zero vector itself is not an element of $\mathcal{D}_0$.

Requiring that credence functions are finite, or non-negative, or positive in all elements, or regular in other ways is artificial in order to aid discussion. Examinations of what happens when these regularity conditions are weakened are always welcome. Probability functions are the strict subset of credence functions for which there exists a measure to which the probability function corresponds with the measure of Ω being 1.

I will concentrate on cases such as the trichotomoy in example 1.1 and restrict my attention to logically coherent credence functions in n-dimensional space, where n is the number of mutually disjoint and collectively exhaustive events. Possible worlds are not credence functions, but they can be embedded by defining the vector elements $\xi_k \in \{0, 1\}$, depending on which of the n events is true, $k = 1, \ldots, n$. This is artificial, especially the choice of the number 1, and any account of epistemic norms must prove itself to be robust if this number is changed to something else that makes sense (see Howson, 2008, 20; for a response see Joyce, 2015; and Pettigrew, 2016, part I, chapter 6).

2.2 Scoring Rules, Entropy, Divergence

Bruno de Finetti shows that the probability functions form the convex hull of possible worlds so embedded (see de Finetti, 2017, subsection 3.4). For any vector c in the vector space of credence functions, there is a vector p in the set of probability functions which is closer to each possible world than c, where closeness is evaluated in terms of metric

distance. If c is not an element of the convex hull of possible worlds, then the vector p is strictly closer to each possible world than c.

There are two important points to make about de Finetti's theorem. (1) It shows in what sense probability functions are privileged over other credence functions. This can be cashed out in terms of pragmatics (see Savage, 1971) or accuracy (see Joyce, 1998), perhaps also in terms of evidence—I am not familiar with the literature. (2) There is a need to define what it means for one credence function to be close to another credence function. One could simply use metric distance. I am hoping to make the case in what follows that this is implausible.

Let us start more modestly with a scoring rule and restrict ourselves to probability functions $\mathcal{P} \subset \mathcal{D}_0$. Probabilism, after all, is not where the geometry of reason and information theory disagree. I have defined scoring rules in equation (1) and the associated strict propriety in definition 1.1. McCarthy characterizes strictly proper scoring rules in a theorem whose proof he omits (see McCarthy, 1956, 654). Thankfully, Arlo Hendrickson and Robert Buehler provide a proof (see Hendrickson and Buehler, 1971, p. 1918).

Definition 2.1. Let M be a convex subset of $\mathcal{D}$, H be a function $H : M \to \mathbb{R}$, and $q, \hat{q} \in M$ such that

$$H(p) \leq \langle p - q, \hat{q} \rangle + H(q) \text{ for all } p \in M. \tag{5}$$

Then $\hat{q}$ is a supergradient of H at q relative to M.

The supergradient is the gradient wherever the function is differentiable.

Definition 2.2. A function $f : V \subset \mathbb{R}^k \to \mathbb{R}$ is homogeneous of degree d if and only if

$$f(\alpha x) = \alpha^d f(x) \text{ for all } \alpha > 0. \tag{6}$$

Use of Euler's homogeneous function theorem (see Zill, 2011) allows for the proof of McCarthy's theorem. Let ∇H be the gradient of the function H if it exists, i.e.

$$\nabla H(x) = \left(\frac{\partial H}{\partial x_1}(x), \ldots, \frac{\partial H}{\partial x_n}(x) \right)^\intercal \tag{7}$$

and

$$\Xi = \{\xi_1, \ldots, \xi_n\}. \tag{8}$$

Theorem 2.1 (McCarthy's Theorem)**.** A scoring rule $S : \Xi \times \mathcal{P} \to \mathbb{R} \cup \{-\infty, \infty\}$ is strictly proper if and only if there exists a function $H : \mathcal{D}_0 \to \mathbb{R}$ which is (a) homogeneous of the first degree, (b) concave, and (c) such that S is a subgradient of H relative to $\mathcal{D}_0$ at p for all $p \in \mathcal{P}$.

It is important to take the partial derivative of H as a function defined on $\mathcal{D}_0$, not just as a function defined on $\mathcal{P}$. As Hendrickson and Buehler point out, this is the error in Marschak, 1959, 97, where the Log score appears to be a counterexample to McCarthy's theorem. None of this is new, but it gives us leverage for what is to follow. Not only is McCarthy's theorem a powerful characterization theorem for strictly proper scoring rules, it also associates an entropy function H and a divergence D with each scoring rule. With McCarthy's result in hand, scoring rules now come as triplets of scoring rules, entropy

functions, and divergences.

Definition 2.3. The entropy H associated with S is defined as per McCarthy's theorem (see theorem 2.1); the divergence associated with S is defined to be

$$D_S(c\|\hat{c}) = H(\hat{c}) - H(c) + \langle c - \hat{c}, S(\hat{c})\rangle. \tag{9}$$

Here are two example, the Log score and the Brier score. All summation indices go from 1 to n. If $x \in \mathcal{P}$, then $\sum_k x_k = 1$. The loss function or scoring rule for the Log score is

$$\text{(LS) } S(\xi_i, x) = \left(\ln \sum_k x_k\right) - \ln x_i. \tag{10}$$

For the Brier score, it is

$$\text{(BS) } S(\xi_i, x) = 1 - \frac{2x_i}{\sum_k x_k} + \sum_j \left(\frac{x_j}{\sum_k x_k}\right)^2. \tag{11}$$

The corresponding entropy functions are

$$\text{(LS) } H(x) = -\sum_i x_i \ln \frac{x_i}{\sum_k x_k} \tag{12}$$

$$\text{(BS) } H(x) = \sum_i x_i \left(1 - \frac{2x_i}{\sum_k x_k} + \sum_j \left(\frac{x_j}{\sum_k x_k}\right)^2\right) \tag{13}$$

The reader can verify that

$$\nabla H(x) = S(x) \tag{14}$$

for both the Log score and the Brier score. This is where we need the entropy to be defined on $\mathcal{D} \supset \mathcal{P}$ in order to avoid Marschak's error from above. Note that the Log score violates LOCALITY on $\mathcal{D} \setminus \mathcal{P}$, so that arguments using the unique characteristic of the Log score to fulfill LOCALITY presupposes an independent argument for probabilism (see Landes, 2015).

The divergence associated with the Log score is

$$D_{\text{LS}}(p\|q) = \sum_i p_i \ln \frac{p_i}{\sum_i p_k} - \sum_i p_i \ln \frac{q_i}{\sum_k q_k}. \tag{15}$$

The divergence associated with the Brier score is

$$D_{\text{BS}}(p\|q) = \sum_i p_i \left[\sum_j \left(\frac{q_j}{\sum_k q_k} - \delta_{ij} \right)^2 - \sum_j \left(\frac{p_j}{\sum_k p_k} - \delta_{ij} \right)^2 \right]. \tag{16}$$

where δ_{ij} is the Kronecker delta. For probability distributions p, q equation (15) is the Kullback-Leibler divergence and equation (16) is the Squared Euclidean Distance.

A strictly proper scoring rule S_1 is a 'close relative' of scoring rule S_2 if the two are positive linear transformations of each other, so

$$S_1(x) = m \cdot S_2(x) + b \tag{17}$$

for some $m \in \mathbb{R}^+$ and $b \in \mathbb{R}^n$. Scoring rules differ from their close relatives only in the sense that they bill or pay in a different currency and provide a different initial penalty or reward. They do not differ from them in terms of optimization (i.e. their extrema are equivalent).

In subsection 3.2, I will show using convex conjugates that only the Brier score (and its close relatives) fulfill SYMMETRY. In subsection 4.3, I will show that only the Log score (and its close relatives) fulfill LOCALITY. These are not new results, although the proof using convex conjugates is original. Once these results are established, I have the tools to address Pettigrew's argument for UNIVOCAL DOMINANCE. In section 5, I show that the Brier score violates REDUCTIO-RESISTANCE and both Brier score and Log score violate HORIZON.

3 Symmetry and Sensitivity

3.1 Symmetry

I will let Pettigrew and Selten speak for themselves as they argue for symmetry. Note that the term symmetry is also used for the feature of a scoring rule to ignore in what order the credences are listed, i.e. for any permutation ρ of $I = \{1, \ldots, n\}$ it is true that

$$S(\xi_i, x) = S(\xi_{\rho(i)}, \hat{x}), \tag{18}$$

where $\hat{x} = (x_{\rho(1)}, \ldots, x_{\rho(n)})^\intercal$. Selten accepts this type of symmetry by arguing that "scores should not depend on the numbering of the alternatives" (see Selten, 1998, 54). For a counterargument see Winkler, 1994. In what follows symmetry means something different: a symmetric scoring rule S is associated with a symmetric divergence D such that

$$D_S(p\|q) = D_S(q\|p) \text{ for all } p \text{ and } q. \tag{19}$$

Selten calls this feature 'neutrality,' Pettigrew calls it symmetry. Pettigrew writes,

> we reason to Symmetry as follows: We have a strong intuition that the inaccuracy of an agent's credence function at a world is the distance between that credence function and the ideal credence function at that world. But we have no strong intuition that this distance must be the distance from the ideal credence function to the agent's credence function rather than the distance to the ideal credence function from the agent's credence function; nor have we a strong intuition that it is the latter rather than the former. But if there were non-symmetric divergences that gave rise to measures of inaccuracy, we would expect that we would have intuitions about this latter question, since, for at least some accounts of the ideal credence function at a world and for some agents, this would make a difference to the inaccuracies to which such a divergence gives rise. Thus, there cannot be such divergences. Symmetry follows. (Pettigrew, 2016, 67f.)

Selten writes about neutrality,

> one looks at the hypothetical case that one and only one of two theories p and q is right, but it is not known which one. The expected score loss of the wrong theory is a measure of how far it is from the truth. It is only fair to require that this measure is 'neutral' in the sense that it treats both theories equally. If p is wrong and q is right, then p should be considered to be as far from the truth as q in the opposite case that q is wrong and p is right. A scoring rule should not be prejudiced in favor of one of both theories in the contest between p and q. The severity of the deviation between them should not be judged differently depending on which of them is true or false. A scoring rule which is not neutral is discriminating on the basis of the location of the theories in the space of all probability distributions over

> the alternatives. Theories in some parts of this space are treated more favorably than those in some other parts without any justification. Therefore, the neutrality axiom is a natural requirement to be imposed on a reasonable scoring rule.

Both Pettigrew and Selten go on to show that the Brier score and its close relatives are the only strictly proper scoring rules fulfilling SYMMETRY, a result already found in Savage, 1971, 788, and one that I will prove in subsection 3.2. My argument against Pettigrew and Selten will become more pronounced in the course of this paper.

Suffice it to say for now that it makes a difference, especially when viewed from the perspective of updating, whether one moves from a distribution p to a distribution q or vice versa (pace Pettigrew). Scoring rules should be partial (and not neutral) in the contest between two theories, when one of them makes much stronger claims than the other (pace Selten). It is the Brier score, after all, which penalizes stronger theories sometimes at the expense of rewarding the less accurate prediction (for an example, see the end of subsection 4.1).

3.2 Symmetry and Brier Score

Theorem 3.1. Only the Brier score and its close relatives fulfill SYMMETRY.

I will provide an original new proof for this theorem (there are more conventional proofs in Savage, 1971, 788; and Selten, 1998, section 4). My proof gives the reader a brief introduction to convex conjugates, which may in many other respects be a fruitful mathematical tool dealing with scoring rules, entropy functions, and divergence functions. Let

$\mathcal{P}$ be the set of probability distributions over a trichotomy-type $n+1$-dimensional outcome space. $x \in \mathcal{P}$ can be represented by the probabilities $x_0, \ldots, x_n$ or by the vector $x = (x_1, \ldots, x_n)^\intercal \in \mathbb{R}^n$ with the usual restrictions on probabilities. If the latter is the case, which will be true for the rest of this subsection, then $x_0 = 1 - \sum x_i$.

This differs in notation from the rest of the paper; and so does my assumption for the rest of the subsection that S is a reward (rather than a loss) function. The reason for the inconsistency of convention is that I will use a fair amount of convex analysis in this subsection. The entropy functions for loss scores are concave. All theorems of convex analysis are valid for concave functions just as much as for convex functions, but in convex analysis, for obvious reasons (to align convex sets with convex functions), the convention is to use convex functions.

The reward function $S_{\text{LS}} : \mathbb{R}^n \to \mathbb{R}^n$ for the Log score is therefore

$$S_{\text{LS}}(x) = (\ln x_i)^\intercal_{i=1,\ldots,n} \tag{20}$$

and the entropy function is

$$H_{\text{LS}}(x) = -\sum_{i=1}^{n} x_i \ln x_i - \left(1 - \sum_{i=1}^{n} x_i\right) \ln \left(1 - \sum_{i=1}^{n} x_i\right). \tag{21}$$

I will use equations (20) and (21) to illustrate convex conjugates.

Let the entropy function H for the scoring rules under consideration be a differentiable, convex function on $\mathcal{P}$.

Definition 3.1. The convex conjugate $H^* : \mathcal{P}^* \to \mathbb{R}$ is defined on

$$\mathcal{P}^* = \{x^* \in \mathbb{R}^n | \nabla H(x) = x^*, x \in \mathcal{P}\} \tag{22}$$

by

$$H^*(x^*) = \sup_{x \in \mathcal{P}} \{\langle x^*, x \rangle - H(x)\}. \tag{23}$$

Convex conjugates induce the so-called Legendre-Fenchel duality between convex differentiable functions on $\mathcal{P}$ and $\mathcal{P}^*$. It can be shown that $H^{**} = H$. Section 12 of R. Tyrrell Rockafellar's book *Convex Analysis* has the details. Rockafellar shows, for example, the following two lemmas (see Rockafellar, 1997, 104f).

Lemma 3.2. Taking conjugates reverses functional inequalities: $f_1 \leq f_2$ implies $f_1^* \geq f_2^*$. A real-valued function f_2 is greater or equal to another real-valued function f_1 on a common domain if on that common domain $f_1(x) \leq f_2(x)$.

Lemma 3.3. $\langle x, x^* \rangle \leq f(x) + f^*(x^*)$ for all $x \in \mathcal{P}, x^* \in \mathcal{P}^*$ and any strictly proper convex function f and its conjugate f^*.

The inequality in lemma 3.3 is called Fenchel's inequality.

What is also of interest to us is that

Lemma 3.4. $(\nabla H)^{-1} = \nabla (H^*)$

and

Lemma 3.5. $D_H(x \| y) = D_{H^*}(y^* \| x^*)$.

Proof. Lemmas 3.4 and 3.5 are proven in Boissonnat et al., 2010, 287. □

My illustration for convex conjugates is the Log score (the Brier score would make a poor illustration, because, as we will find out, the Brier score is uniquely self-dual). For a given $x \in \mathcal{P}$, differentiate (21) to define x^*

$$x_i^* = \frac{\partial}{\partial x_i} H_{\text{LS}}(x) = \ln\left(1 - \sum_{j=1}^{n} x_j\right) - \ln x_i. \tag{24}$$

To find ∇H_{LS}^*, which is equal to $(\nabla H_{\text{LS}})^{-1}$ according to lemma 3.4, solve the system of equations

$$\sum_{j=1}^{n} \left(e^{x_i^*} + \delta_{ij}\right) x_j = 1. \tag{25}$$

It has the solution

$$x_i = \frac{1 + \left(\sum_{i=1}^{n} e^{x_i^*}\right) - n e^{x_i^*}}{1 + \left(\sum_{i=1}^{n} e^{x_i^*}\right)}. \tag{26}$$

Because of lemma 3.4, just as $\nabla H_{\text{LS}}(x)$ corresponds to the right-hand side of (24), $\nabla H_{\text{LS}}^*(x^*)$ corresponds to the right-hand side of (26). H_{LS}^* is an antiderivative of the partial derivatives in (26) with the constant of integration fixed by definition 3.1.

Lemma 3.6. The only self-dual convex differentiable function on $\mathcal{P}$ is $F(x) = \frac{1}{2}\langle x, x\rangle$.

Proof. I am adapting a proof in Rockafellar, 1997, 106. That F is self-dual follows from definition 3.1 and lemma 3.4. Now let G be a self-dual convex differentiable function. By

lemma 3.3 (Fenchel's inequality),

$$\langle x, x \rangle \leq G(x) + G^*(x) = 2G(x). \tag{27}$$

This means that $G \geq F$. By lemma 3.2, $G^* \leq F^*$. Together with the self-duality of both F and G, this implies $F = G$. □

Lemmas 3.4, 3.5, and 3.6 establish theorem 3.1.

3.3 Sensitivity

Selten attributes two undesirable features of a scoring rule to the Log score: lack of sensitivity and hypersensitivity. The two features are best explained by example.

Example 3.1 (Zero Dark Thirty)**.** In the movie *Zero Dark Thirty* (written by Mark Boal), the character Maya interrupts the deputy director of the CIA to deliver the following probability assessment of whether or not Osama bin Laden is in a particular location in Pakistan in early 2011 (OBL): "A hundred percent he's there. Okay, fine, ninety-five percent because I know how certainty freaks you guys out, but it's a hundred." Maya is uncertain about other things: whether or not, for example, Dominique Strauss-Kahn will win the next presidential election in France (DSK); or whether Arnold Schwarzenegger and Maria Shriver will still be married in 2012 (AMS). Let Maya's forecast over these events (OBL, DSK, AMS) be based on $(1.00, 0.38, 0.91)$ and Maya's commitment to coherence in terms of logical entailment and probabilism.

Maya's forecast is insensitive with respect to the Log score in the sense that no matter

what she believes about DSK and AMS, if OBL is false her penalty will be infinite. Unless one indulges the regularity fetish of 'you guys' at the CIA, this may not be an uncommon scenario. What we traditionally call knowledge may in some ways be about ruling things out probabilistically. If those forecasts are predicted frequencies of repeatable experiments, a sufficient amount of data will set the penalty at infinity, no matter how good these forecasts are with respect to the non-extreme probabilities that they contain. The Log score is insensitive to them, while the Brier score is not.

At the end of subsection 5.4.5, I describe a scenario where one way for an advocate of the Brier score to weasel out of a dilemma described there is to embrace regularity. The Log score's option to avoid violating SENSITIVITY is also to embrace regularity. I am using the term regularity here in Jeffrey's sense: if the event space is finite, then nothing but logical contradictions and tautologies is assigned extreme probability. I will use the term REGULARITY in subsection 5.2 in a more narrow but not unrelated sense.

Selten goes on to charge the Log score with hypersensitivity. Here is, again, an example.

Example 3.2 (Zero Dark Thirty Again)**.** Rowan and Jessie are CIA agents with a regularity fetish who, however, largely agree with Maya's confident assessment. Rowan sets the probability for not-OBL at 1×10^{-2}, Jessie at 1×10^{-6}. Rowan's distribution diverges in terms of the Log score from Jessie's by 0.08216, almost three times as much as the divergence of Taylor's forecast from Quinn's. Taylor and Quinn are two other CIA analysts, who report forecasts for not-OBL of 20% and 30% respectively (divergence is 0.0282). Their forecasts are 10 percentage points apart, whereas Rowan's and Jessie's are less than 1 percentage point apart.

Selten argues that small differences between small probabilities ought not to be taken this seriously and considers the hypersensitivity feature a "very undesirable one" (see Selten, 1998, 51). Similar to a debate in confirmation theory, the war of intuitions is between difference and ratio. Selten's argument only succeeds if the target audience is already inclined towards measures of difference. After all, Rowan's forecast for non-OBL is 10,000 times greater than Jessie's, while Quinn's is not even twice that of Taylor's. In conclusion, an advocate of the Log score can address Selten's sensitivity argument by recourse to regularity; and reject Selten's hypersensitivity argument by pointing out that it preaches to the choir.

4 Locality

4.1 Rewarding Uncertainty About Non-Realized Outcomes

The Brier score, the Spherical score (another strictly proper scoring rule), and many other scoring rules depend on all components of the vector p representing a probabilistic credence function. A scoring rule fulfilling the LOCALITY requirement only depends on the probability assigned to the event that is the realized outcome (for a characterization of local scoring rules that are local in a less restrictive sense see Dawid et al., 2012). In subsection 4.3, I show (using Leonard J. Savage's proof) that the only scoring rules fulfilling LOCALITY are the Log score and its not relevantly different close relatives.

Example 4.1 (Tokens)**.** Before Casey draws from a bag with n kinds of tokens in it (colour 1, colour 2, ..., colour n), Tatum reports the forecast $(p_1, \ldots, p_n)$ of associated

credences. Tatum's forecast agrees with the axioms of probability.

Let p_1 be fixed and colour 1 be the realized outcome. If the Brier score is used, Tatum's penalty T depends on $p_2, \ldots, p_{n-1}$ and is

$$T(p_2, \ldots, p_{n-1}) = 1 - 2p_1 + \sum_{i=1}^{n-1} p_i^2 + \left(1 - \sum_{i=1}^{n-1} p_i\right)^2. \tag{28}$$

T reaches its minimum where $p_i = \frac{1-p_1}{n-1}$ for $i = 2, \ldots, n-1$. The higher the entropy of Tatum's non-realized probabilities, the less stinging Tatum's penalty. The Brier score thus penalizes Tatum (1) for not correctly identifying colour 1 as the realized outcome, but also (2) for reporting variation in the non-realized probabilities. Even though this is the Brier score, it has a ring of information theory to it. The Log score depends only on the realized probability. I.J. Good appears to have favoured such a scoring rule (see Good, 1952, 112).

Because I feel the intuitive appeal of information theory, I consider LOCALITY to be only weakly plausible. There is a sense in which you may want to reward a forecaster not only for assigning a high probability to the realized outcome, but also for uncertainty about the outcomes that were not realized. Of course, doing so sometimes results in a greater loss for Tatum than for Casey even if Tatum assigned a higher probability to the realized outcome. As an example, let Tatum's forecast be $(0.12, 0.86, 0.02)$ and Casey's be $(0.10, 0.54, 0.36)$. Even though Tatum assigned 12% to colour 1 while Casey assigned 10%, and Casey drew a token of colour 1, Tatum is penalized more severely at 1.5144 compared to Casey at 1.2312 using the Brier score. The greater penalty for Tatum may strike one as counterintuitive, but it is a natural consequence of a scoring rule violating LOCALITY.

4.2 Bronfman Objection

Here is how LOCALITY may still work in favour of information theory against the geometry of reason. To set the scene, I should mention that there are at least two publication anomalies in the study of scoring rules. I already mentioned the earlier one: McCarthy omitted the proof to one of its most important theorems. The proof, using Euler's Theorem, is not trivial and was published almost twenty years later by Hendrickson and Buehler. The later publication anomaly is that Aaron Bronfman wrote an excellent article about a problem with using supervaluationist semantics to justify probabilism (see Bronfman, 2009). Then he decided not to publish it. The manuscript has circulated and is available online. It is called "A Gap in Joyce's Argument for Probabilism."

Consider the result (in Predd et al., 2009) that vindicates probabilism non-pragmatically by demonstrating that given a particular continuous and strictly proper scoring rule, any non-probabilistic credence c is accuracy-dominated by a probabilistic credence function p (depending on c) in the sense that p is strictly closer to all possible worlds (and therefore more accurate) than c. No probabilistic credence function is accuracy-dominated in this way (see also Joyce, 1998). The proof is inspired by de Finetti's theorem referred to in subsection 2.2.

I owe the following characterization of Bronfman's objection to Pettigrew (see chapter 5 in Pettigrew, 2016; for the Pater Peperium case see Paul, 2016).

Example 4.2 (Pater Peperium)**.** I must choose between three sandwich options: Marmite, cheese, and Pater Peperium (or Gentleman's Relish).

I have eaten cheese sandwiches before and feel indifferent about them. I have never had

a Marmite or Pater Peperium sandwich, but know that people either love Marmite and hate Pater Peperium or vice versa. There appears to be nothing irrational about choosing the cheese sandwich even though either way (whether I am of the love-marmite-hate-pater-peperium or hate-marmite-love-pater-peperium type) there is a better sandwich to choose.

Joyce has shown that for any strictly proper scoring rule, a non-probabilistic credence function is accuracy dominated by a probabilistic credence function. The Bronfman objection is that you can show that there is always another strictly proper scoring rule (Bronfman shows that only having two candidate quadratic loss scoring rules suffices to make this point) by which moving from the accuracy dominated credence function to the probabilistic credence function results in a loss at some possible world. Unless we settle on a unique scoring rule to do the accounting, Joyce's non-pragmatic vindication of probabilism is undermined.

Pettigrew uses Bronfman's objection to propose UNIVOCAL DOMINANCE. It is an appealing feature of a scoring rule to have some claim to uniqueness in order to address Bronfman's objection. The Brier score has this claim: it is (up to linear transformation, which do not make a relevant difference) the only strictly proper scoring rule which fulfills SYMMETRY. Unfortunately for Pettigrew, the Log score also has a claim to uniqueness. It is the only strictly proper scoring rule which fulfills LOCALITY. We could now haggle over which uniqueness claim is stronger. In some ways, this paper is meant to undermine the intuitive appeal of SYMMETRY altogether. I will not, however, push UNIVOCAL DOMINANCE and LOCALITY as joint justification for the Log score, as Pettigrew pushes UNIVOCAL DOMINANCE and SYMMETRY as joint justification for the Brier score.

Pettigrew's argument is suspect (he by no means is unaware of its tenuous appeal and reiterates that many of his results stand even if UNIVOCAL DOMINANCE is implausible). Let a uniqueness claim have dependent and independent reasons. The dependent reasons justify the uniqueness on account of the unique features that the object of the uniqueness claim exhibits. The independent reasons make no reference to these features, but provide a reason to have a successful candidate for winning the uniqueness contest. I do not see how these independent reasons add to the epistemic justification for the uniqueness claim.

When Räuber Hotzenplotz puts a loaded pistol on your chest and asks you which of your three children is your favourite child and you name one of them, then there may be uniqueness features about this child that identify him or her as your favourite. Even the fact that you named this child may be one of those dependent reasons, but the independent reason that Hotzenplotz pressed you for the identification does not epistemically count towards making it more plausible that this child is your favourite child or that indeed you have a favourite child.

4.3 Proof of Locality Uniqueness for the Log Score

Let there be a function $f : (a, b) \to \mathbb{R}$ ($a, b \in \mathbb{R}$ with $a < b$) for which the following is true: for every $x \in (a, b)$ there exists a linear function $L_x : \mathbb{R} \to \mathbb{R}$ such that $L_x(x) = x$ and

$$L_x(y) \leq f(y) \text{ for all } y \in (a, b) \tag{29}$$

with inequality whenever $y \neq x$. The conditions basically say that f has a subgradient at every point in its domain. Then the function is strictly convex, i.e.

$$f(\lambda y + (1-\lambda)\hat{y}) < \lambda f(y) + (1-\lambda) f(\hat{y}) \tag{30}$$

for all $0 < \lambda < 1$ and for all $y, \hat{y}$ in the domain of f. A theorem of convex analysis tells us that

Lemma 4.1. f is almost everywhere differentiable (i.e. the set of points where it is not differentiable is countable).

Proof. See theorem 25.5 in Rockafellar, 1997, 246. □

Consider two real-valued differentiable functions F and G whose domain is $(0, 1)$.

Lemma 4.2. If $xyF(xy) = (1-x)yG((1-x)y)$ for all $x, y \in (0, 1)$, then $zF(z) = zG(z) = k$ for all $z \in (0, 1)$ and some constant $k \in \mathbb{R}$.

Proof. $xyF(xy) = (1-x)yG((1-x)y)$ and $(1-x)yF((1-x)y) = xyG(xy)$ implies that $xy(F-G)(xy) = (x-1)(F-G)((1-x)y)$. For $x = 0.5$ this means that $(F-G)(0.5y) = 0$, and since y is arbitrary, this implies $F = G$ on $(0, 1)$. Therefore, we have $xyF(xy) = (1-x)yF((1-x)y)$. Rewrite as

$$\frac{F(xy)}{F((1-x)y)} = \frac{1-x}{x}. \tag{31}$$

For any $s, t \in (0, 1)$ with $s + t < 1$, we can take $s = xy, t = (1-x)y$ where $y = s + t$ and

$x = s/(s+t)$. Then

$$\frac{F(s)}{F(t)} = \frac{1-x}{x} = \frac{t}{s}, \tag{32}$$

which means that $zF(z) = k$ for some constant $k \in \mathbb{R}$. □

Lemma 4.3. A function $f : (0,1) \to \mathbb{R}$ which fulfills $xf'(x) = k$ for all $x \in (0,1)$ and some $k \in \mathbb{R}$ is of the form $f(x) = k \ln x + b$ for some $b \in \mathbb{R}$.

Proof. Integrate $f'(x) = k/x$. □

Define $f_i(p_i) = S(\xi_i, p_i)$ for a scoring rule fulfilling LOCALITY and PROPRIETY. f_i are functions from $[0,1] \to \mathbb{R} \cup \{-\infty, \infty\}, i = 1, \ldots, n$. Strict propriety (see definition 1.1) requires

$$\sum_{i=1}^{n} p_i f_i(p_i) \geq \sum_{i=1}^{n} q_i f_i(p_i) \text{ for all } p, q \in \mathcal{P} \tag{33}$$

with inequality whenever $p \neq q$; $p = (p_1, \ldots, p_n)^\intercal$ and $q = (q_1, \ldots, q_n)^\intercal$. Let $0 < k < 1$ be arbitrary, but fixed, and define $g_k : (0,1) \to \mathbb{R}$

$$g_k(x) = xf_1(x \cdot k) + (1-x)f_2((1-x) \cdot k). \tag{34}$$

Let $\alpha, x \in (0,1)$. Define the linear function

$$L_{x,k}(\alpha) = \alpha f_1(x \cdot k) + (1-\alpha)f_2((1-x) \cdot k) \tag{35}$$

and the two distributions

$$\begin{aligned} p &= (\alpha \cdot k, (1-\alpha) \cdot k, \tfrac{1-k}{n-2}, \ldots, \tfrac{1-k}{n-2}) \\ q &= (x \cdot k, (1-x) \cdot k, \tfrac{1-k}{n-2}, \ldots, \tfrac{1-k}{n-2}) \end{aligned}$$

Then (33) implies

$$g_k(\alpha) \geq L_{x,k}(\alpha) \text{ for all } \alpha \in (0,1) \tag{36}$$

with inequality whenever $\alpha \neq x$ and $g_k(x) = L_{x,k}(x)$.

As a consequence of lemma 4.1, g_k is strictly convex and differentiable almost everywhere. At points $x \in (0,1)$ where g_k is differentiable, $L_{x,k}$ is the tangent line and therefore

$$g_k'(x) = f_1(xk) - f_2((1-x)k), \tag{37}$$

whereas differentiating (34) yields

$$g_k'(x) = \left[f_1(xk) - f_2((1-x)k)\right] + k \cdot \left[x f_1'(xk) - (1-x) f_2'((1-x)k)\right]. \tag{38}$$

(37) and (38) equal each other, therefore

$$xk f_1'(xk) = (1-x)k f_2'((1-x)k). \tag{39}$$

Use lemma 4.2, lemma 4.3, and (39) for the proof of the following theorem. Note, however, that the Log score only fulfills LOCALITY if the domain on which LOCALITY is evaluated

is probabilistic credences. Jürgen Landes discusses how to save some notion of locality for logarithmic scoring rules which are strictly proper for all credence functions (see Landes, 2015).

Theorem 4.4. The only strictly proper scoring rules which fulfill LOCALITY are the Log score and its close relatives.

5 Updating

5.1 Information Theory and the Geometry of Reason

Joyce uses a norm of gradational accuracy together with six axioms (structure, extensionality, dominance, normality, weak convexity, symmetry) to give an epistemic justification of probabilism: the requirement for a rational agent to keep her partial beliefs in keeping with the axioms of probability theory.

It is a natural question to ask whether the same line of reasoning can give us an epistemic justification of standard conditioning and Jeffrey conditioning. The former updates partial beliefs in light of an event that is known to be true based on intervening evidence, i.e. $P'(E) = 1$ (P' for the posterior). The latter updates partial beliefs in light of an event about which the agent has shifted in uncertainty based on intervening evidence, i.e. $P'(E) = y$ where y is not necessarily equal to $P(E) = x$, the prior of E. In a series of articles that will be pivotal for the rest of this paper, Hannes Leitgeb and Richard Pettigrew show that using Joyce's approach, accuracy can be bifurcated into local and global accuracy. For this bifurcation to give consistent results, the Brier score must be used for Joyce's norm of

gradational accuracy. Given Leitgeb and Pettigrew's assumptions, standard conditioning is vindicated, but Jeffrey conditioning is ruled out. 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 (based on the Brier score) fails a host of expectations that are reasonable to have for the kind of updating scenario that LP conditioning addresses. Therefore, the Brier score fails REDUCTIO-RESISTANCE. 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 (this geometric relationship is not necessarily Euclidean). At first, this assumption is natural: probability distributions can be isomorphically identified with points in a simplex. It seems natural to measure the difference between probability distributions by applying the Euclidean metric defined on the larger space containing the simplex. 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.

Leitgeb and Pettigrew consider 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).

For information theory, as opposed to the geometry of reason, the underlying topology for credence functions is not a metric space (see figures 1 and 2 for illustration). The

term information geometry is due to Imre Csiszár, who considers the Kullback-Leibler divergence a non-commutative (asymmetric) analogue of squared Euclidean distance and derives several results that are intuitive information geometric counterparts of standard results in Euclidean geometry (see chapter 3 of Csiszár and Shields, 2004).

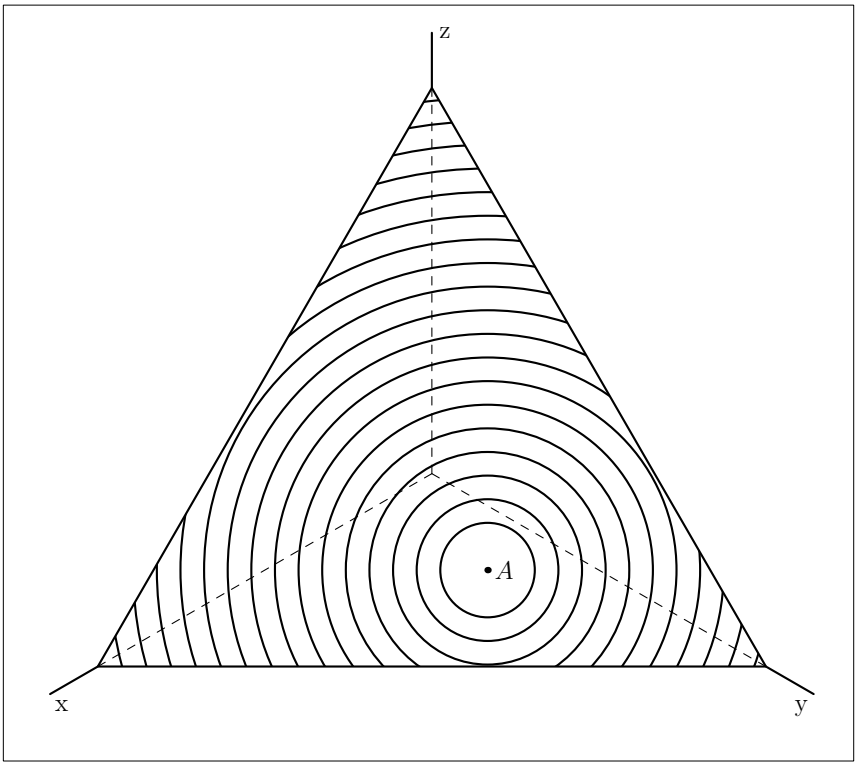

Figure 1: The simplex $\mathbb{S}^2$ in three-dimensional space $\mathbb{R}^3$ with contour lines corresponding to the geometry of reason around point A in equation (40). Points on the same contour line are equidistant from A with respect to the Euclidean metric. Compare the contour lines here to figure 2. Note that this diagram and all the following diagrams are frontal views of the simplex.

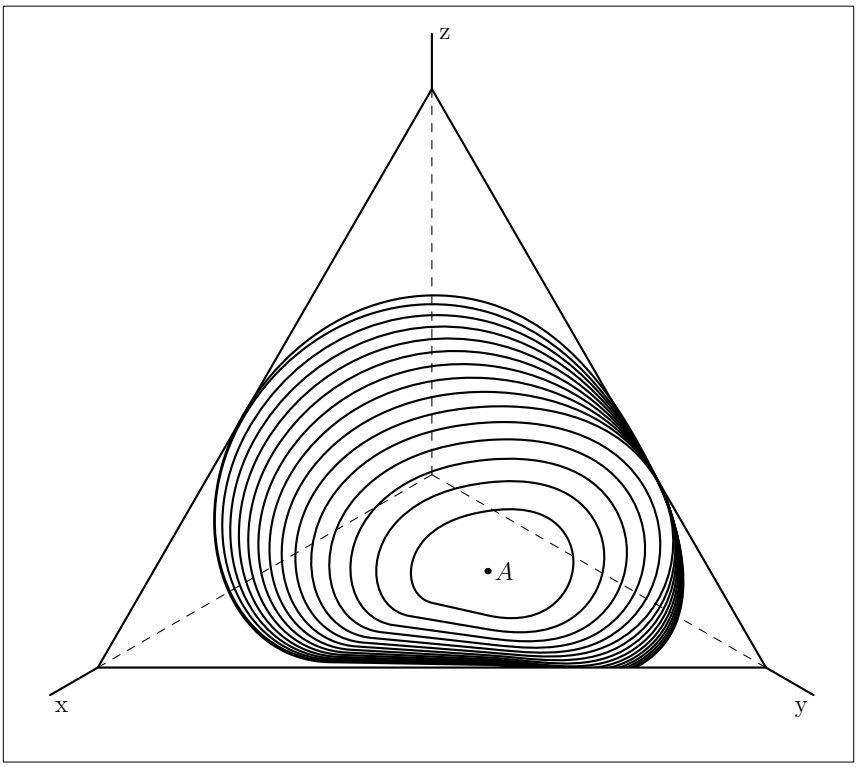

Figure 2: The simplex $\mathbb{S}^2$ with contour lines corresponding to information theory around point A in equation (40). Points on the same contour line are equidistant from A with respect to the Kullback-Leibler divergence. The contrast to figure 1 will become clear in much more detail in the body 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.

5.2 Jeffrey-Type Updating Scenarios

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

Example 5.1 (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 5.1 in the next section. For now note that LP conditioning violates all of the following plausible expectations for updating methods in Jeffrey-type updating scenarios. This is **List A**:

- CONTINUITY An updating method ought to be continuous with standard conditioning as a limiting case.
- REGULARITY An updating method 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 updating method ought not to give "extremely unattractive" results in a Levinstein scenario (see Levinstein, 2012, where Benjamin Levinstein not only

articulates this failed expectation for LP conditioning, but also the previous two; the following three are original to this article).

- INVARIANCE An updating method ought to be partition invariant in the sense that introducing irrelevant subpartitioning does not change the outcome of the update.

- EXPANSIBILITY An updating method ought to be insensitive to an expansion of the event space by zero-probability events.

- HORIZON An updating method 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 updating method 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 original (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.

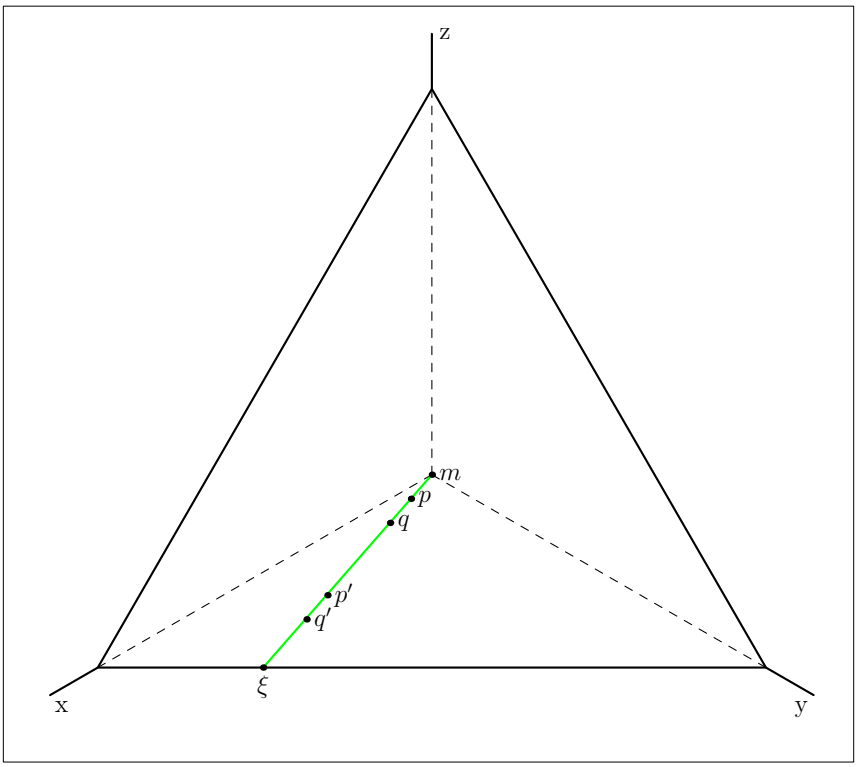

Figure 3: An illustrations of conditions (i)–(iii) for COLLINEAR HORIZON in **List B**. p, p' and q, q' must be equidistant and collinear with m and ξ. If q, q' is more peripheral than p, p', then COLLINEAR HORIZON requires that $|d(p, p')| < |d(q, q')|$.

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

Information theory has its own axiomatic approach to justifying standard conditioning (see Shore and Johnson, 1980); it also provides a justification for Jeffrey conditioning and

generalizes it (see Lukits, 2015). All of these virtues stand in contrast to the violations of the expectations in **List B**. The rest of this paper fills in the details of these violations both for the geometry of reason and 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.

5.3 LP Versus Jeffrey Conditioning

Here is a simple example corresponding to example 5.1 (Sherlock Holmes) 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:

$$a = \left(\frac{1}{3}, \frac{1}{2}, \frac{1}{6}\right) \qquad b = \left(\frac{1}{2}, \frac{3}{8}, \frac{1}{8}\right) \qquad c = \left(\frac{1}{2}, \frac{5}{12}, \frac{1}{12}\right). \tag{40}$$

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 $D_{\text{KL}}(B, A)$ 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):

$$D_{\mathrm{KL}}(B, A) = \sum_{i=1}^{3} b_i \log \frac{b_i}{a_i}. \tag{41}$$

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:

$$\|B - A\| = \sqrt{\sum_{i=1}^{n} (b_i - a_i)^2}. \tag{42}$$

What is remarkable about the three points in (40) is that

$$\|C - A\| \approx 0.204 < \|B - A\| \approx 0.212 \tag{43}$$

and

$$D_{\mathrm{KL}}(B, A) \approx 0.0589 < D_{\mathrm{KL}}(C, A) \approx 0.069. \tag{44}$$

The Kullback-Leibler divergence and Euclidean distance give different recommendations

with respect to proximity. In terms of the global inaccuracy measure presented in Leitgeb and Pettigrew (see Leitgeb and Pettigrew, 2010a, 206) and $E = W$ (all possible worlds are epistemically accessible),

$$\mathrm{GExp}_A(C) \approx 0.653 < \mathrm{GExp}_A(B) \approx 0.656. \tag{45}$$

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 5.1 about Sherlock Holmes) and nothing stronger, then information theory via the Kullback-Leibler divergence recommends the posterior corresponding to B; the geometry of reason as expounded in Leitgeb and Pettigrew recommends the posterior corresponding to C.

Here is a brief account of Jeffrey conditioning and LP conditioning, which are competing methods to arrive at posterior probability distributions in Jeffrey-type updating scenarios. They roughly align with information theory and the geometry of reason, respectively, in the context of this paper. There are, however, many formal epistemologists who defend the epistemic norm of Jeffrey conditioning but are skeptical about further claims of information theory, such as updating in scenarios that pose an affine constraint which are not Jeffrey-type updating scenarios, for example Bas van Fraassen's Judy Benjamin problem (see van Fraassen, 1981; for the debate over information theory and affine constraints see Lukits, 2014).

Example 5.2 (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 5.2, let

$$\begin{aligned} P(E_1) &= 1/3 \\ P(E_2) &= 1/2 \\ P(E_3) &= 1/6 \end{aligned} \tag{46}$$

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 Richard 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,

$$\begin{aligned} P'_{\text{JC}}(E_i) &= P'(E_i|F^*)P'(F^*) + P'(E_i|F^{**})P'(F^{**}) \\ &= P(E_i|F^*)P'(F^*) + P(E_i|F^{**})P'(F^{**}) \end{aligned} \tag{47}$$

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, given their assumptions. 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 (LP conditioning). It proceeds as follows for example 5.2 and in general provides the minimally inaccurate posterior probability distribution in Jeffrey-type updating scenarios according to the geometry of reason.

Solve the following two equations for x and y:

$$\begin{aligned} P(E_1) + x &= P'(F^*) \\ P(E_2) + y + P(E_3) + y &= P'(F^{**}) \end{aligned} \tag{48}$$

and then set

$$\begin{aligned} P'_{\text{LP}}(E_1) &= P(E_1) + x \\ P'_{\text{LP}}(E_2) &= P(E_2) + y \\ P'_{\text{LP}}(E_3) &= P(E_3) + y \end{aligned} \tag{49}$$

For the more formal and more general account see Leitgeb and Pettigrew, 2010b, 254. The results for example 5.2 are:

$$\begin{aligned} P'_{\text{LP}}(E_1) &= 1/2 \\ P'_{\text{LP}}(E_2) &= 5/12 \\ P'_{\text{LP}}(E_3) &= 1/12 \end{aligned} \tag{50}$$

Compare these results to the results of Jeffrey conditioning:

$$\begin{aligned} P'_{\text{JC}}(E_1) &= 1/2 \\ P'_{\text{JC}}(E_2) &= 3/8 \\ P'_{\text{JC}}(E_3) &= 1/8 \end{aligned} \tag{51}$$

Note that (46), (51), and (50) correspond to A, B, C in (40) (see figure 4).

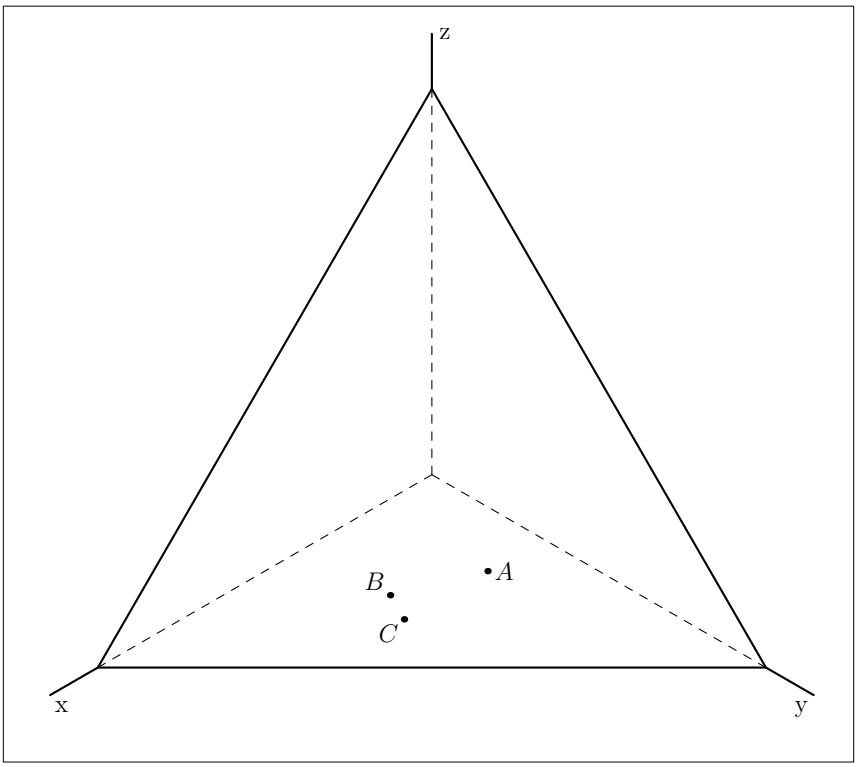

Figure 4: The simplex $\mathbb{S}^2$ in three-dimensional space $\mathbb{R}^3$ with points a, b, c as in equation (40) representing probability distributions A, B, C. Note that geometrically speaking C is closer to A than B is. Using the Kullback-Leibler divergence, however, B is closer to A than C is.

5.4 Expectation Violations for the Geometry of Reason

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

5.4.1 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 fulfilling CONDITIONING (the requirement in subsection 2.1) is to Leitgeb and Pettigrew.

To illustrate a CONTINUITY violation, consider the case where Sherlock Holmes reduces his credence that the culprit was male to $\varepsilon_n = 1/n$ for $n = 4, 5, \ldots$. The sequence ε_n is not meant to reflect a case where Sherlock Holmes becomes successively more certain that the culprit was female. It is meant to reflect countably many parallel scenarios which only differ by the degree to which Sherlock Holmes is sure that the culprit was female. These parallel scenarios give rise to a parallel sequence (as opposed to a successive sequence) of updated probabilities $P'_{\text{LP}}(F^{**})$ and another sequence of updated probabilities $P'_{\text{JC}}(F^{**})$ (F^{**} is the proposition that the culprit is female). As $n \to \infty$, both of these sequences go to one.

Straightforward conditionalization on the evidence that 'the culprit is female' gives us

$$\begin{aligned} P'_{\text{SC}}(E_1) &= 0 \\ P'_{\text{SC}}(E_2) &= 3/4 \\ P'_{\text{SC}}(E_3) &= 1/4 \end{aligned} \tag{52}$$

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

$$\begin{aligned} P'_{\text{JC}}(E_1) &= 1/n \rightarrow 0 \\ P'_{\text{JC}}(E_2) &= 3(n-1)/4n \rightarrow 3/4 \\ P'_{\text{JC}}(E_3) &= (n-1)/4n \rightarrow 1/4 \end{aligned} \tag{53}$$

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

$$\begin{aligned} P'_{\text{LP}}(E_1) &= 1/n \rightarrow 0 \\ P'_{\text{LP}}(E_2) &= (4n-3)/6n \rightarrow 2/3 \\ P'_{\text{LP}}(E_3) &= (2n-5)/6n \rightarrow 1/3 \end{aligned} \tag{54}$$

LP conditioning violates CONTINUITY.

5.4.2 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 (see Jeffrey, 1987).

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 5.3 (Regularity Holmes)**.** Everything is as in example 5.1, 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 (46).

Then his posterior probabilities look as follows:

$$\begin{aligned} P'_{\text{JC}}(E_1) &= 2/3 \\ P'_{\text{JC}}(E_2) &= 1/4 \\ P'_{\text{JC}}(E_3) &= 1/12 \end{aligned} \tag{55}$$

$$\begin{aligned} P'_{\text{LP}}(E_1) &= 2/3 \\ P'_{\text{LP}}(E_2) &= 1/3 \\ P'_{\text{LP}}(E_3) &= 0 \end{aligned} \tag{56}$$

With LP conditioning, Sherlock Holmes' subjective probability that Ms. T is the culprit in example 5.3 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.

5.4.3 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 Levinstein's example:

Example 5.4 (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.

5.4.4 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. In the following example, Sherlock Holmes and Jane Marple agree on all relevant facts and on their prior probabilities, but LP conditioning leads to a divergence in posterior probabilities.

Example 5.5 (Jane Marple)**.** Jane Marple is on the same case as Sherlock Holmes in example 5.1 and arrives at the same relatively prior probability distribution as Sherlock Holmes (I will call Jane Marple's relatively prior probability distribution Q and her posterior probability distribution Q'). Jane Marple, however, has a more finely grained probability assignment than Sherlock Holmes and distinguishes between the case where Ms. S went to boarding school with her, of which she has a vague memory, and the case where Ms. S did not and the vague memory is only about a fleeting resemblance of Ms. S with another boarding school mate. Whether or not Ms. S went to boarding school with Jane Marple is completely beside the point with respect to the crime, and Jane Marple considers the possibilities equiprobable whether or not Ms. S went to boarding school with her.

Let $E_2 \equiv E_2^* \vee E_2^{**}$, where E_2^* is the proposition that Ms. S is the culprit and she went to boarding school with Jane Marple and E_2^{**} is the proposition that Ms. S is the culprit and she did not go to boarding school with Jane Marple. Then

$$\begin{aligned} Q(E_1) &= 1/3 \\ Q(E_2^*) &= 1/4 \\ Q(E_2^{**}) &= 1/4 \\ Q(E_3) &= 1/6 \end{aligned} \tag{57}$$

Now note that while Sherlock Holmes and Jane Marple agree on the relevant facts of the criminal case (who is the culprit?) in their posterior probabilities if they use Jeffrey conditioning,

$$\begin{aligned} P'_{\text{JC}}(E_1) &= 1/2 \\ P'_{\text{JC}}(E_2) &= 3/8 \\ P'_{\text{JC}}(E_3) &= 1/8 \end{aligned} \tag{58}$$

$$\begin{aligned} Q'_{\text{JC}}(E_1) &= 1/2 \\ Q'_{\text{JC}}(E_2^*) &= 3/16 \\ Q'_{\text{JC}}(E_2^{**}) &= 3/16 \\ Q'_{\text{JC}}(E_3) &= 1/8 \end{aligned} \tag{59}$$

they do not agree if they use LP conditioning,

$$\begin{aligned} P'_{\text{LP}}(E_1) &= 1/2 \\ P'_{\text{LP}}(E_2) &= 5/12 \\ P'_{\text{LP}}(E_3) &= 1/12 \end{aligned} \tag{60}$$

$$\begin{aligned} Q'_{\text{LP}}(E_1) &= 1/2 \\ Q'_{\text{LP}}(E_2^*) &= 7/36 \\ Q'_{\text{LP}}(E_2^{**}) &= 7/36 \\ Q'_{\text{LP}}(E_3) &= 1/9 \end{aligned} \tag{61}$$

LP conditioning violates INVARIANCE.

5.4.5 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

$$\begin{aligned} P(X_1) &= 0 \\ P(X_2) &= 0.3 \\ P(X_3) &= 0.6 \\ P(X_4) &= 0.1 \end{aligned} \tag{62}$$

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 (62) or on the following list:

$$\begin{aligned} P(X_2) &= 0.3 \\ P(X_3) &= 0.6 \\ P(X_4) &= 0.1 \end{aligned} \tag{63}$$

Whether you update on (62) or (63) 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 the geometry of reason is joined at the hip with LP conditioning which runs afoul of REGULARITY even when it updates on priors and evidence that obey REGULARITY (see subsection 5.4.2).

LP conditioning violates EXPANSIBILITY. Selten has a requirement for scoring rules similar to EXPANSIBILITY that he calls elongation invariance (axiom 2 in Selten, 1998, 54). It is to some degree ironic that Selten uses elongation invariance to show that the Brier score and its close relatives are the only reasonable scoring rules, when a decade later Leitgeb and Pettigrew derive an updating method from the geometry of reason so intimately associated with the Brier score that violates an immediate analogy to Selten's elongation invariance.

5.4.6 Horizon

Example 5.6 (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 5.7 (Airplane Gliders)**.** Compare two scenarios. In the first, an airplane which is considered safe (probability of crashing is $1/10^9$) goes through an inspection where a mechanical problem is found which increases the probability of a crash to 1/100. In the second, military gliders land behind enemy lines, where their risk of perishing is 26%. A slight change in weather pattern increases this risk to 27%. (Schlesinger, 1995, 211.)

I claim that an updating method ought to fulfill the requirements of the horizon effect: it ought to be more difficult to update as probabilities become more extreme (or less

middling). I have formalized this requirement in **List B**. It is trivial that the geometry of reason does not fulfill it. Information theory fails as well, which gives the horizon effect its prominent place in both lists. The way information theory fails, however, is quite different. Near the boundary of $\mathbb{S}^{n-1}$, information theory reflects the horizon effect just as our expectation requires. The problem is near the centre, where some equidistant points are more divergent the closer they are to the middle. I will give an example and more explanation in subsection 5.5.2.

5.5 Expectations for Information Theory

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. Going in the opposite direction, however, from $P(X) = 0$ to $P'(X) = x' > 0$ is controversial and unusual.

Associated with the Log score via McCarthy's theorem (theorem 2.1) is 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. It 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.

5.5.1 Triangularity

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

$$d(A, C) = d(A, B) + d(B, C), \tag{64}$$

where d is the difference measure we are using, so $d(A, B) = \|B - A\|$ for the geometry of reason and $d(A, B) = D_{\text{KL}}(B, A)$ 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. Call B a zero-sum point between A and C if (64) holds true. For the geometry of reason, the zero-sum points are simply the points on the straight line between A and C. For information geometry, the zero-sum points are the boundary points of the set where you can take a shortcut by making a detour, i.e. all points for which $d(A, B) + d(B, C) < d(A, C)$.

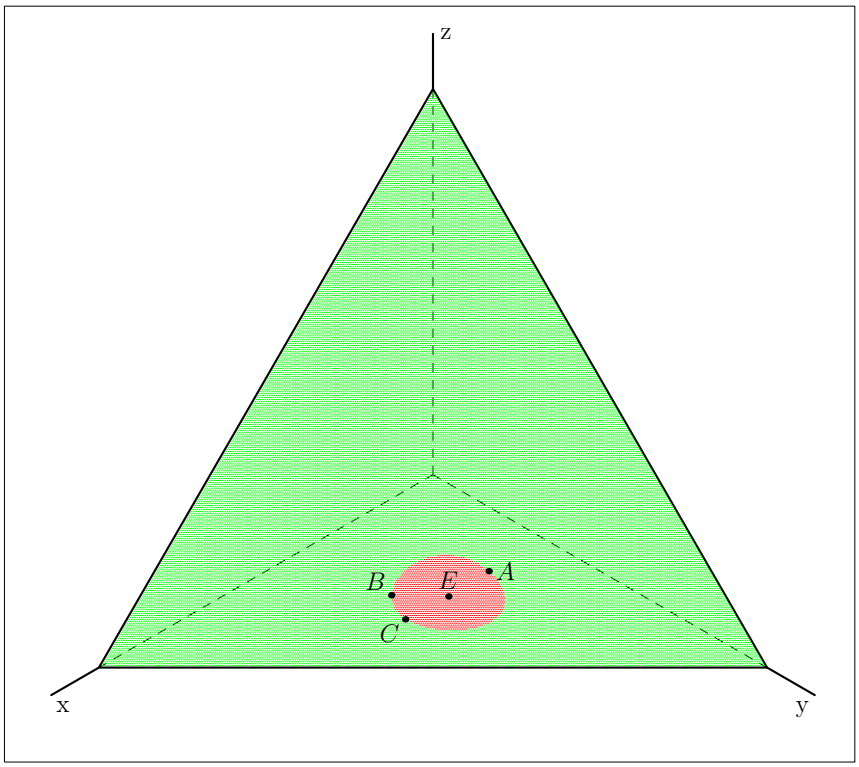

Figure 5: The zero-sum line between A and C is the boundary line between the green area, where the triangle inequality holds, and the red area, where the triangle inequality is violated. The posterior probability distribution B recommended by Jeffrey conditioning always lies on the zero-sum line between the prior A and the LP posterior C, as per equation (65). E is the point in the red area where the triangle inequality is most efficiently violated. Even though it can be calculated using the Lambert W function, $e_k = \frac{c_k}{W\left(\frac{c_k}{a_k}\exp(1+\lambda)\right)}$, with λ chosen to fulfill $\sum e_k = 1$, it is not clear to me whether E is the midpoint between A and C or not.

Proposition 5.1. 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,

$$D_{\mathrm{KL}}(C, A) = D_{\mathrm{KL}}(B, A) + D_{\mathrm{KL}}(C, B). \tag{65}$$

Proof. To prove equation (65) 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; and point recommended by LP conditioning) can be expressed using three variables suitably constrained to yield probabilities (for example, $\alpha - \beta > 0$):

$$\begin{aligned} A &= (1-\alpha, \beta, \alpha-\beta) \\ B &= \left(1-\gamma, \tfrac{\gamma\beta}{\alpha}, \tfrac{\gamma(\alpha-\beta)}{\alpha}\right) \\ C &= \left(1-\gamma, \beta+\tfrac{1}{2}(\gamma-\alpha), \alpha-\beta+\tfrac{1}{2}(\gamma-\alpha)\right) \end{aligned} \tag{66}$$

The rest is basic algebra using the definition of the Kullback-Leibler divergence in (41).

$$(1-\gamma)\log\frac{1-\gamma}{1-\alpha} + \left(\beta + \frac{1}{2}(\gamma-\alpha)\right)\log\frac{\beta+\frac{1}{2}(\gamma-\alpha)}{\beta} +$$

$$\left(\alpha-\beta+\frac{1}{2}(\gamma-\alpha)\right)\log\frac{\alpha-\beta+\frac{1}{2}(\gamma-\alpha)}{\alpha-\beta} =$$

$$(1-\gamma)\log\frac{1-\gamma}{1-\alpha}+\frac{\gamma\beta}{\alpha}\log\frac{\frac{\gamma\beta}{\alpha}}{\beta}+\frac{\gamma(\alpha-\beta)}{\alpha}\log\frac{\frac{\gamma(\alpha-\beta)}{\alpha}}{\alpha-\beta}+$$

$$(1-\gamma)\log\frac{1-\gamma}{1-\gamma}+\left(\beta+\frac{1}{2}(\gamma-\alpha)\right)\log\frac{\beta+\frac{1}{2}(\gamma-\alpha)}{\frac{\gamma\beta}{\alpha}}+$$

$$\left(\alpha-\beta+\frac{1}{2}(\gamma-\alpha)\right)\log\frac{\alpha-\beta+\frac{1}{2}(\gamma-\alpha)}{\frac{\gamma(\alpha-\beta)}{\alpha}} \tag{67}$$

In order to prove the claim for arbitrary n one simply generalizes (66). □

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

Before we get carried away with these analogies between divergences and metrics, however, it is important to note that it is not appropriate to impose expectations that are conventional for metrics on divergences. Bregman divergences, for example, in some sense violate the triangle equality by design. If d_H is a Bregman divergence with the corresponding

convex entropy function H, then for a convex set $\mathcal{C} \in \mathbb{R}^n$ and all $x \in \mathcal{C}$ and $y \in \mathbb{R}^n$ the following reverse triangle inequality is true:

$$d_H(x, y) \geq d_H(x, y') + d_H(y', y), \tag{68}$$

where y' is the projection of y onto $\mathcal{C}$ such that $d_H(y', y) = \min\{d_H(z, y), z \in \mathcal{C}\}$. The squared Euclidean distance is an interesting case in point for this property. In a generalization of the Pythagorean theorem, $c^2 > a^2 + b^2$ holds for obtuse triangles. When $\mathcal{C}$ is affine (such as a plane in $\mathbb{R}^3$), (68) turns from an inequality to an equation (replacing '$\geq$' by '=') for all Bregman divergences. For the squared Euclidean distance, (68) is then the conventional Pythagorean theorem. To subject the difference concept between probability distributions to a TRIANGULARITY requirement may be a temptation to resist and only reveal another instance of the Euclidean prejudice identified by Mormann.

The three points A, B, C in (40) violate TRIANGULARITY for D_{KL} because

$$0.067806 = D_{\mathrm{KL}}(A, B) + D_{\mathrm{KL}}(B, C) < D_{\mathrm{KL}}(A, C) = 0.071530. \tag{69}$$

Information theory, however, does not only violate TRIANGULARITY. It violates it in a particularly egregious way.

Proposition 5.2. Let x and z be distinct points on $\mathbb{S}^{n-1}$ with coordinates x_i and z_i $(1 \leq i \leq n)$. 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:

$$D_{\mathrm{KL}}(z, x) > D_{\mathrm{KL}}(y, x) + D_{\mathrm{KL}}(z, y). \tag{70}$$

Proof. It is straightforward to see that (70) is equivalent to

$$\sum_{i=1}^{n}(z_i - x_i)\log\frac{\vartheta x_i + (1-\vartheta)z_i}{x_i} > 0. \tag{71}$$

Expand the right hand side to

$$\sum_{i=1}^{n}\left(z_i + \frac{\vartheta}{1-\vartheta}x_i - \frac{\vartheta}{1-\vartheta}x_i - x_i\right)\log\frac{\frac{1}{1-\vartheta}\left(\vartheta x_i + (1-\vartheta)z_i\right)}{\frac{1}{1-\vartheta}x_i} > 0. \tag{72}$$

(72) is clearly equivalent to (71). It is also equivalent to

$$\sum_{i=1}^{n}\left(z_i + \frac{\vartheta}{1-\vartheta}x_i\right)\log\frac{z_i + \frac{\vartheta}{1-\vartheta}x_i}{\frac{1}{1-\vartheta}x_i} + \sum_{i=1}^{n}\frac{1}{1-\vartheta}x_i\log\frac{\frac{1}{1-\vartheta}x_i}{z_i + \frac{\vartheta}{1-\vartheta}x_i} > 0, \tag{73}$$

which is true by Gibbs' inequality. □

Like Bregman divergences in general, the Kullback-Leibler divergence in particular violates TRIANGULARITY by design. Giving proposition 5.2 a misguided and paradoxical reading from the intuitions of geometry, the more often you stop on the way, the faster you reach your destination.

5.5.2 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 example 5.7 (Airplane Gliders) make a persuasive case for curvature.

Information theory is characterized by an intuitively opaque '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 our examples. It is not trivial, however, to articulate what the horizon requirement really demands.

COLLINEAR HORIZON in **List B** seeks to set up the requirement as weakly as possible, only demanding that points collinear with the centre exhibit the horizon effect. The hope is that continuity will take care of the rest, since the horizon effect is desirable also for probability distributions that are not collinear with the centre. The way I have formalized the HORIZON and COLLINEAR HORIZON requirement is artificial in the face of the more comprehensive epistemic intuition. COLLINEAR HORIZON especially is dependent on the Euclidean idea of collinearity. In a more integrated account it would be desirable to have these requirements reformulated in a more general fashion; convexity may play a major role in such a reformulation. It could, perhaps, identify a divergence that does not violate HORIZON or reveal why the divergence of information theory violates it.

The Kullback-Leibler divergence fails COLLINEAR HORIZON. Here is a simple example.

$$p = \left(\frac{1}{5}, \frac{2}{5}, \frac{2}{5}\right) \qquad p' = q = \left(\frac{1}{4}, \frac{3}{8}, \frac{3}{8}\right) \qquad q' = \left(\frac{3}{10}, \frac{7}{20}, \frac{7}{20}\right) \tag{74}$$

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 $\|p' - p\| = \|q' - q\|$ and m, p, p', q, q' are

collinear. In violation of COLLINEAR HORIZON,

$$D_{\mathrm{KL}}(p, p') = 7.3820 \cdot 10^{-3} > 6.4015 \cdot 10^{-3} = D_{\mathrm{KL}}(q, q'). \tag{75}$$

Just as there is still a reasonable disagreement about difference measures (which do not exhibit the horizon effect) and ratio measures (which do) in confirmation theory, most of us will not have strong intuitions about the adequacy of information theory based on its violation of COLLINEAR HORIZON. One way in which I can attenuate the independent appeal of this violation against information theory is by making it parasitic on the asymmetry of information theory.

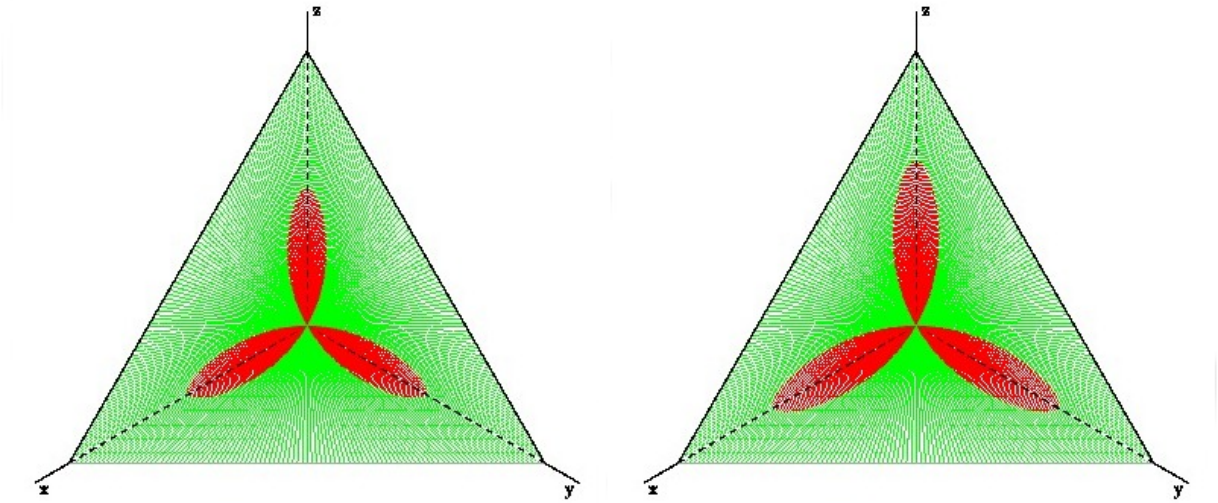

Figure 6: These two diagrams illustrate inequalities (77) and (78). The former displays all points in red which violate COLLINEAR HORIZON, measured from the centre. The latter displays points in different colours whose orientation of asymmetry differs, measured from the centre. The two red sets are not the same, but there appears to be a relationship, one that ultimately I suspect to be due to the more basic property of asymmetry.

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

$$\text{(i) } D_{\text{KL}}(Y, M) > D_{\text{KL}}(M, Y) \text{ and (ii) } D_{\text{KL}}(X, M) > D_{\text{KL}}(Y, X) \tag{76}$$

In terms of coordinates, the inequalities reduce to (H being the Shannon entropy)

$$\text{(i) } H(y) < \frac{1}{n}\sum(\log y_i) - \log\frac{1}{n^2} \text{ and} \tag{77}$$

$$\text{(ii) } H(y) > \log\frac{4}{n} - \sum\left[\left(\frac{3}{2}y_i + \frac{1}{2n}\right)\log\left(y_i + \frac{1}{n}\right)\right]. \tag{78}$$

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

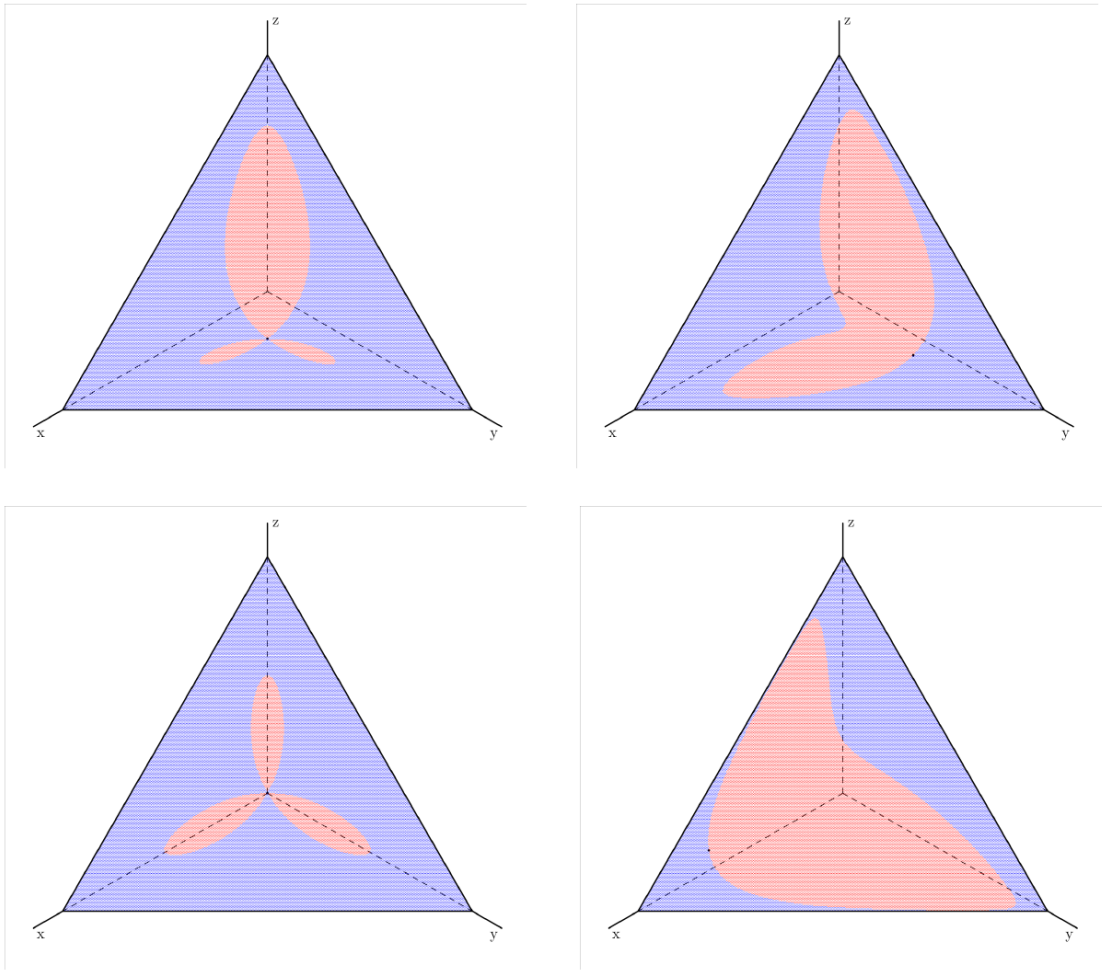

Figure 7: The partition (80) based on different values for P. From top left to bottom right, $P = (0.4, 0.4, 0.2); P = (0.242, 0.604, 0.154); P = (1/3, 1/3, 1/3); P = (0.741, 0.087, 0.172)$. Note that for the geometry of reason, the diagrams are trivial. The challenge for information theory is to explain the non-triviality of these diagrams epistemically without begging the question.

It is not transparent what motivates information theory not only to put probability distributions farther apart near the periphery, as expected, but also near the centre. I lack the epistemic intuition reflected in this behaviour. The next subsection on asymmetry deals with this lack of epistemic intuition writ large.

5.5.3 Transitivity of Asymmetry

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

Extreme probabilities are special and create asymmetries in updating: moving in direction from certainty to uncertainty is asymmetrical to moving in direction from uncertainty to certainty. Geometry of reason's metric topology, however, allows for no asymmetries.

Example 5.8 (Extreme Asymmetry)**.** Consider two cases where for case 1 the prior probabilities are $Y_1 = (0.4, 0.3, 0.3)$ and the posterior probabilities are $Y_1' = (0, 0.5, 0.5)$; for case 2 the prior probabilities are reversed, so $Y_2 = (0, 0.5, 0.5)$ and the posterior probabilities $Y_2' = (0.4, 0.3, 0.3)$.

Case 1 is a straightforward application of standard conditioning. Case 2 is more complicated: what does it take to raise a prior probability of zero to a positive number? In terms of information theory, the information required is infinite. Case 2 is also not compatible with standard conditioning (at least not with what Alan Hájek calls the ratio analysis of

conditional probability, see Hájek, 2003). The geometry of reason may want to solve this problem by signing on to a version of regularity but is ill-advised to do so because LP conditioning violates REGULARITY independently. Continuity, a consistent view of regularity, and symmetry: the geometry of reason cannot consistently hold all three.

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.

The simplex $\mathbb{S}^2$ represents all the probability distributions for trichotomy. 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

$$\Delta_P(P') = D_{\mathrm{KL}}(P', P) - D_{\mathrm{KL}}(P, P'), \tag{79}$$

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

$$\Delta^{-1}(\mathbb{R}_{>0}) \qquad \Delta^{-1}(\mathbb{R}_{<0}) \qquad \Delta^{-1}(\{0\}). \tag{80}$$

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 (80) is non-trivial—some probability distributions are of the Q-type, some are of the R-type, and it is difficult to think of an epistemic distinction between them that does not already presuppose information theory (see figure 7 for illustration). The Legendre-Fenchel duality from subsection 3.2 may be helpful, or else a more coordinate-free approach using differential geometry. I do not know.

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) $d(x, x) = 0$ (self-similarity)

(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 B**. For an example of it, consider

$$P_1 = \left(\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\right) \qquad P_2 = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right) \qquad P_3 = \left(\frac{2}{5}, \frac{2}{5}, \frac{1}{5}\right). \tag{81}$$

In the terminology of TRANSITIVITY OF ASYMMETRY in **List B**, (P_1, P_2) is asymmetrically positive, and so is (P_2, P_3). The reasonable expectation is that (P_1, P_3) is asymmetrically positive by transitivity, but for the example in (81) it is asymmetrically negative.

How counterintuitive this is (epistemically and otherwise) is demonstrated by the fact that in MDS (the multi-dimensional scaling of distance relationships) almost all asymmetric distance relationships under consideration are asymmetrically transitive in this sense, for examples see international trade in Chino, 1978; journal citation in Coombs, 1964; car switch in Harshman et al., 1982; telephone calls in Harshman and Lundy, 1984; interaction or input-output flow in migration, economic activity, and social mobility in Coxon, 1982; flight time between two cities in Gentleman et al., 2006, 191; mutual intelligibility between Swedish and Danish in van Ommen et al., 2013, 193; Tobler's wind model in Tobler, 1975; and the cyclist lovingly hand-sketched in Kopperman, 1988, 91.

This 'ill behaviour' of information theory begs for explanation. 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.

6 Conclusion

I have shown that the account provided by defenders of the geometry of reason, such as Leitgeb, Pettigrew, or Selten, is indefensible. While it is true that the scoring rule associated with the geometry of reason, the Brier score, is unique in fulfilling a symmetry requirement, I have shown both conceptually and by example that asymmetry is preferable because it allows for greater sensitivity to extreme probabilities. The Log score associated with information theory fulfills the requirement to be sensitive to extremity and therefore asymmetric. However, the Log score behaves oddly toward the centre of probability distributions (for example, its asymmetry is not transitive) and is in general an ill-behaved measure of divergence (for example, it not only violates triangularity but does so in egregious ways as demonstrated in the paper).

Both the geometry of reason and information theory are well-established and highly integrated theories, with many interesting results and various ways in which they have intuitive appeal. The geometry of reason has in its favour that it makes use of our geometric intuition as well as the substantial mathematical apparatus that comes along with

it. In a table in subsection 2.1 I have summarized, however, that the geometry of reason fails several plausible requirements (INFORMATION, LOCALITY, HORIZON, REDUCTIO-RESISTANCE) and uniquely fulfills only implausible requirements (GEOMETRY, SYMMETRY).

A requirement that Pettigrew lists in favour of the geometry of reason, UNIVOCAL DOMINANCE, is fulfilled via SYMMETRY, but information theory also fulfills it via LOCALITY, and the requirement itself is suspect (see subsection 4.2). Most significantly, besides the violation of sensitivity to extremity, the geometry of reason does not agree with intuitions that most of us have about updating. At the same time, this is one of the major strengths of information theory. Both standard conditioning (Bayesian conditioning) and Jeffrey conditioning can be justified using information theory. Information theory licences an updating method on affine constraints, the principle of minimum cross-entropy, which generalizes both standard and Jeffrey conditioning.

Information theory also gives us the entropy function we have come to expect on the basis of Shannon's analysis of entropy in Shannon, 1948. It fulfills Shannon's axioms, while the entropy function associated with the geometry of reason fails them. The work, however, is not complete. Further investigations may reveal why information theory saddles us with a divergence function as ill-behaved as the Kullback-Leibler divergence. I have introduced the Legendre-Fenchel duality in subsection 3.2, which has potential to shed some light on this issue.

More promising yet is the use of differential geometry in the formal theory of partial beliefs. This paper is still entrenched in thinking of partial beliefs as parametrized by probabilities $p_i, i = 1, \ldots, n$. The geometry of reason is dedicated to this entrenchment on

principle. Information theory can hopefully escape it. There are many other parameters that uniquely identify probability distributions, for example

$$\left(\ln \frac{p_1}{p_n}, \ldots, \ln \frac{p_{n-1}}{p_n}\right)^{\intercal}. \tag{82}$$

Once a categorical distribution (corresponding to finite outcome spaces) is parametrized this way, it joins the legion of other distributions in the exponential family (normal, exponential, gamma, chi-squared, beta, Dirichlet, Bernoulli, Poisson, Wishart, inverse Wishart, geometric) and partakes in substantial results for these distributions. The geometry of reason prevents us from pursuing these fruitful abstractions by tying us to our geometric intuitions. Whether information theory proves to be useful is a question for which I eagerly await answers.

References

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

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

Boissonnat, Jean-Daniel, Frank Nielsen, and Richard Nock. "Bregman Voronoi Diagrams." *Discrete and Computational Geometry* 44, 2: (2010) 281–307.

Bregman, Lev. "The Relaxation Method of Finding the Common Point of Convex Sets

and Its Application to the Solution of Problems in Convex Programming." *USSR Computational Mathematics and Mathematical Physics* 7, 3: (1967) 200–217.

Bronfman, Aaron. "A Gap In Joyce's Argument for Probabilism.", 2009. University of Michigan: unpublished manuscript.

Chino, Naohito. "A Graphical Technique for Representing the Asymmetric Relationships Between N Objects." *Behaviormetrika* 5, 5: (1978) 23–44.

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

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

Csiszár, Imre, and Paul C Shields. *Information Theory and Statistics: A Tutorial*. Hanover, MA: Now Publishers, 2004.

Dawid, A Philip, Steffen Lauritzen, and Matthew Parry. "Proper Local Scoring Rules on Discrete Sample Spaces." *Annals of Statistics* 40, 1: (2012) 593–608.

De Finetti, Bruno. *Theory of Probability*. Chichester, UK: Wiley, 2017.

Diaconis, Persi, and Sandy Zabell. "Updating Subjective Probability." *Journal of the American Statistical Association* 77, 380: (1982) 822–830.

van Fraassen, Bas. "A Problem for Relative Information Minimizers in Probability Kinematics." *British Journal for the Philosophy of Science* 32, 4: (1981) 375–379.

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, I.J. "Rational Decisions." *Journal of the Royal Statistical Society. Series B (Methodological)* 14, 1: (1952) 107–114.

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

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.

Hendrickson, Arlo, and Robert Buehler. "Proper Scores for Probability Forecasters." *Annals of Mathematical Statistics* 42, 6: (1971) 1916–1921.

Howson, Colin. "De Finetti, Countable Additivity, Consistency and Coherence." *British Journal for the Philosophy of Science* 59, 1: (2008) 1–23.

Howson, Colin, and Allan Franklin. "Bayesian Conditionalization and Probability Kinematics." *British Journal for the Philosophy of Science* 45, 2: (1994) 451–466.

Jeffrey, Richard. *The Logic of Decision*. New York, NY: McGraw-Hill, 1965.

———. "Alias Smith and Jones: The Testimony of the Senses." *Erkenntnis* 26, 3: (1987) 391–399.

Joyce, James. "A Nonpragmatic Vindication of Probabilism." *Philosophy of Science* 65, 4: (1998) 575–603.

———. "The Value of Truth: A Reply to Howson." *Analysis* 75, 3: (2015) 413–424.

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

Landes, Jürgen. "Probabilism, Entropies and Strictly Proper Scoring Rules." *International Journal of Approximate Reasoning* 63: (2015) 1–21.

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* 191, 7: (2014) 1409–1431.

———. "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.

Marschak, Jacob. "Remarks on the Economics of Information." Technical report, Cowles Foundation for Research in Economics, Yale University, 1959.

McCarthy, John. "Measures of the Value of Information." *Proceedings of the National Academy of Sciences* 42, 9: (1956) 654–655.

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.

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

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.

Paul, L.A. *Transformative Experience.* Oxford, UK: Oxford University, 2016.

Pettigrew, Richard. *Accuracy and the Laws of Credence.* Oxford, UK: Oxford University, 2016.

Predd, Joel, Robert Seiringer, Elliott Lieb, Daniel Osherson, H. Vincent Poor, and Sanjeev Kulkarni. "Probabilistic Coherence and Proper Scoring Rules." *IEEE Transactions on Information Theory* 55, 10: (2009) 4786–4792.

Rockafellar, Ralph. *Convex Analysis.* Princeton, N.J: Princeton University, 1997.

Savage, Leonard. "Elicitation of Personal Probabilities and Expectations." *Journal of the American Statistical Association* 66, 336: (1971) 783–801.

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

Selten, Reinhard. "Axiomatic Characterization of the Quadratic Scoring Rule." *Experimental Economics* 1, 1: (1998) 43–62.

Shannon, Claude. "A Mathematical Theory of Communication." *Bell System Technical Journal* 27, 1: (1948) 379–423, 623–656.

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.

Tobler, Waldo. *Spatial Interaction Patterns.* Schloss Laxenburg, Austria: International Institute for Applied Systems Analysis, 1975.

Wagner, Carl. "Probability Kinematics and Commutativity." *Philosophy of Science* 69, 2: (2002) 266–278.

Winkler, Robert. "Evaluating Probabilities: Asymmetric Scoring Rules." *Management Science* 40, 11: (1994) 1395–1405.

Zill, Dennis. *Multivariable Calculus.* Sudbury, MA: Jones and Bartlett, 2011.