European Journal for Philosophy of Science (2021) 11:79
https://doi.org/10.1007/s13194-021-00398-x

PAPER IN FORMAL METHODS AND EXACT SCIENCES

Symmetry and partial belief geometry

Stefan Lukits[1]

Received: 18 June 2020 / Accepted: 6 July 2021 / Published online: 29 July 2021

Abstract

When beliefs are quantified as credences, they are related to each other in terms of closeness and accuracy. The "accuracy first" approach in formal epistemology wants to establish a normative account for credences (probabilism, Bayesian conditioning, principle of indifference, and so on) based entirely on the alethic properties of the credence: how close it is to the truth. To pull off this project, there is a need for a scoring rule. There is widespread agreement about some constraints on this scoring rule (for example propriety), but not whether a unique scoring rule stands above the rest. The Brier score equips credences with a structure similar to metric space and induces a "geometry of reason." It enjoys great popularity in the current debate. I point out many of its weaknesses in this article. An alternative approach 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; and speculate on its remediation.

Keywords Bayesian conditionalization · Epistemic utility · Formal epistemology · Geometry of reason · Gradational accuracy · Information theory · Jeffrey conditioning · Partial beliefs; principle of maximum entropy · Probabilism · Probability kinematics · Probability update · Convex analysis · Scoring rules · Probabilistic forecast

This article belongs to the Topical Collection: EPSA2019: Selected papers from the biennial conference in Geneva
Guest Editors: Anouk Barberousse, Richard Dawid, Marcel Weber

✉ Stefan Lukits
st.lukits@utoronto.ca

[1] Symmetry and Partial Belief Geometry, University of Toronto, Toronto, Canada

1 Introduction

There is a problem with the canonical quantitative representation of partial beliefs in formal epistemology. This representation takes its cues primarily from probability theory where real numbers are assigned to events following the axioms for a probability measure. Formal epistemology would profit by moving from its current orientation toward probability theory to differential geometry, where a partial belief is represented by an element of a differential manifold. The well-established mathematical theory of differential geometry provides a coordinate-free framework that avoids and explains some of the problems with a "coordinated" approach that I reveal in this paper.

Specifying the alternative framework for quantitative partial belief representation in differential geometry is a research project that I am pursuing elsewhere. The focus of this paper is to demonstrate the shortcomings of the coordinated approach currently in use in the research of formal epistemologists. These shortcomings become more apparent with certain symmetry assumptions to which formal epistemologists also frequently appeal. The symmetry assumptions are often associated with scoring rules for partial beliefs that exclusively reward and penalize on epistemic grounds. These scoring rules will be of particular interest to me, as they have been to much of the most recent research activity in this field.

I am restricting myself to finite algebras of propositions. Let some mutually exclusive and collectively exhaustive outcomes (or possible worlds) be $\xi_1, \ldots, \xi_n$. Let the agent's "report" over these n outcomes be $c = (c_1, \ldots, c_n)^\intercal$. These are the coordinates of what I call the coordinated approach. In the coordinate-free approach of differential geometry, they are replaced by an element of a differential manifold which can be locally mapped to an open subset of n-dimensional Euclidean space.

The agent's report may represent a partial belief or an attempt at prediction. The transpose $^\intercal$ symbol merely turns the list of numbers into a column 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-dimensional vector space.

The agent receives a penalty for reporting c according to a loss function. An epistemic agent wants to minimize the penalty without caring 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. The codomain of S is $\mathbb{R} \cup \{\infty\}$. $S(c)$ 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 a random process selects the outcome modeling the event in question.

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

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

Strict propriety narrows down the set of acceptable scoring rules. John McCarthy shows that it requires the existence of a concave entropy function, for which the scoring rule is a derivative of sorts (see McCarthy, 1956). These scoring rules are associated with a particular type of divergence called Bregman divergence (see Bregman, 1967).

Are there 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 Bregman divergence 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 as superior to asymmetric ones (for example in Selten, 1998; Pettigrew, 2016, 80).

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 accessible to intuition by analogy to n-dimensonal Euclidean space. 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 & Pettigrew, 2010, 210), the geometry of reason. Its associated scoring rule is the Brier score. All symmetric scoring rules can with trivial modifications be reduced to the Brier score.

In the tradition of a Schopenhauerian critique directed at Kant's symmetries (see the appendix to volume I of *The World as Will and Representation*), 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 ... 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" (see Mormann (2005), 433–435; also Miller (1984); Jorge Luis Borges echoes Schopenhauer's complaint that in Kant's critique "everything is sacrificed to a rage for symmetry," see Borges (1962), 55).

Information theory, like the geometry of reason, has an associated scoring rule: the Log score. The Log score is asymmetric, but there are other asymmetric strictly proper scoring rules. Pettigrew has an 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.

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. However, I will go further and show how counterintuitive implications of information theory are if we insist on quantitative partial belief representation in terms of coordinates. The problem is not with the Brier score, but with the coordinated approach. If it is not jettisoned in favour of a coordinate-free alternative, I find myself in the Absurdistan outlined in the paper, even with the Log score of information theory. Fortunately, McCarthy's theorem and the entropy function following from it provides enough structure to render a differential geometry approach hopeful in terms of inferring rationality requirements (probabilism, conditioning, etc.).

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 at a model that generates a 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.

littoral compression The divergence function associated with the SR has a tendency to measure distributions near the high-entropy centre as being closer together than distributions near the low-entropy extremes, all else being equal. This requirement depends on coordinated quantitative belief representation, and I will seek to lead it ad absurdum.

ratio conservation When a credence is updated, rational agents are required by certain epistemic norms (for example, standard conditioning) to conserve ratios of prior probabilities for posterior probabilities as long as the events in question are not affected by the evidence on which the agent performs the update. A scoring rule may conform to ratio conservation by recommending updated credences which conserve ratios.

cardinality independence When a credence is updated, rational agents are required to update independently of the cardinality of outcome sets. I can illustrate this most easily with probabilities: if you initially consider a die fair, where each of the sides is rolled with probability 1/6, it should not make a difference to your update if you add an event "rolling a seven" to your outcome space and assign probability zero to it. In the paper I show that with weak assumptions CARDINALITY DEPENDENCE is equivalent to RATIO CONSERVATION.

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

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 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 the outcome space Ω being 1.

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 $\mathcal{D}_0$.

I will 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). A coordinate-free approach no longer depends on such arbitrary choices, as it can filter them out using quotient spaces.

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 a suitable measure of closeness, for example a continuous strictly proper scoring rule (Predd et al. (2009), call continuous strictly proper scoring rules "proper scoring rules" and use the corresponding Definition 2 to prove de Finetti's result in Theorem 1 on page 4788). If c is not a probability function, then the vector p is strictly closer to each possible world than c. If c is a probability function, then one trivially chooses $p = c$.

Probabilism is not where the geometry of reason and information theory disagree, so for the moment I will only look at probability functions $\mathcal{P} \subset \mathcal{D}_0$. I have defined scoring rules in Eq. 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 the proof, see Theorem 2.1.

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 k if and only if

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

Theorem 2.1 (Euler's Homogeneous Function Theorem) *Let $f : \mathbb{R}^n \to \mathbb{R}$ be homogeneous of degree k with continuous first-order partial derivatives. Then*

$$k \cdot H = \sum_{i=1}^{n} x_i \frac{\partial f}{\partial x_i}. \tag{7}$$

Proof The proof is widely available online (see also Pemberton & Rau, 2016, 284). □

Let ∇H be the gradient of the function H if it exists, i.e.

$$\nabla H(x) = \left(\frac{\partial H}{\partial x_1}(x), \ldots, \frac{\partial H}{\partial x_n}(x) \right)^{\mathsf{T}} \tag{8}$$

and

$$\Xi = \{\xi_1, \ldots, \xi_n\}. \tag{9}$$

Theorem 2.2 (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 $p \in \mathcal{P}$.*

Proof The proof is in Hendrickson and Buehler (1971), page 1918, and uses Euler's Homogeneous Function Theorem. Note that whether a function is concave or convex is without deeper significance. An epistemic loss function gives us a concave entropy function while an epistemic reward function gives us a convex entropy function. What matters is the convexity of sets (rather than functions); a convex function has a convex epigraph while a concave function has a convex hypograph. □

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

$$D_S(c\|\hat{c}) = H(\hat{c}) - H(c) + \langle c - \hat{c}, S(\hat{c})\rangle. \tag{10}$$

Corollary 2.1 *As a corollary to Euler's theorem, it is true for a differentiable entropy function H of a strictly proper scoring rule S that*

$$H(x) = \sum_{i=1}^{n} x_i S(\xi_i, x) \tag{11}$$

Corollary 2.1 is useful in determining the entropy function based on a given scoring rule.

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 follows. 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. I hope to show in future research that this provides enough structure for a coordinate-free representation theorem (with the admittedly strong assumption of differentiability for credence functions).

Here are two examples, 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

$$\text{(LS) } S(\xi_i, x) = \left(\ln \sum_k x_k\right) - \ln x_i. \tag{12}$$

For the Brier score, it is

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

The corresponding entropy functions are

$$\text{(LS) } H(x) = -\sum_i x_i \ln \frac{x_i}{\sum_k x_k} \tag{14}$$

$$\text{(BS) } H(x) = \sum_i x_i \left(1 - \frac{2x_i}{\sum_k x_k} + \sum_j \left(\frac{x_j}{\sum_k x_k}\right)^2\right) \tag{15}$$

The reader can verify that the gradient of H equals S,

$$\nabla H(x) = S(x) \tag{16}$$

for both the Log score and the Brier score. This is where we need the entropy to be defined on $\mathcal{D}_0 \supset \mathcal{P}$ 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

$$D_{\mathrm{LS}}(p\|q) = \sum_i p_i \ln \frac{p_i}{\sum_i p_k} - \sum_i p_i \ln \frac{q_i}{\sum_k q_k}. \tag{17}$$

The divergence associated with the Brier score is

$$D_{\mathrm{BS}}(p\|q) = \sum_i p_i \left[\sum_j \left(\frac{q_j}{\sum_k q_k} - \delta_{ij} \right)^2 - \sum_j \left(\frac{p_j}{\sum_k p_k} - \delta_{ij} \right)^2 \right]. \tag{18}$$

where δ_{ij} is the Kronecker delta. For probability distributions p, q (17) is the Kullback-Leibler divergence and Eq. 18 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

$$S_1(x) = m \cdot S_2(x) + b \tag{19}$$

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 trade in a different currency and provide a different initial penalty or reward. They do not differ from them in terms of optimization (their extrema are equivalent).

Only the Brier score (and its close relatives) fulfill SYMMETRY. Only the Log score (and its close relatives) fulfill LOCALITY. Once these results are established, I have the tools to address Pettigrew's argument for UNIVOCAL DOMINANCE. First a few words about RATIO CONSERVATION and CARDINALITY INDEPENDENCE, which the Brier score violates and the Log score fulfills.

3 Ratio conservation

Here is an example that discriminates between the scoring rules in interesting ways.

Example 3.1 (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?

Consider the following three points in three-dimensional space:

$$a = \left(\frac{1}{3}, \frac{1}{2}, \frac{1}{6}\right) b = \left(\frac{1}{2}, \frac{3}{8}, \frac{1}{8}\right) c = \left(\frac{1}{2}, \frac{5}{12}, \frac{1}{12}\right). \tag{20}$$

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

three mutually exclusive events. Now call $D_{\mathrm{KL}}(B, A)$ the Kullback-Leibler divergence of B from A defined as follows, where a_i are the Cartesian coordinates of a (the base of the logarithm is not important, in order to facilitate easy differentiation I will use the natural logarithm):

$$D_{\mathrm{KL}}(B, A) = \sum_{i=1}^{3} b_i \log \frac{b_i}{a_i}. \tag{21}$$

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:

$$\|B - A\| = \sqrt{\sum_{i=1}^{n} (b_i - a_i)^2}. \tag{22}$$

What is remarkable about the three points in Eq. 20 is that

$$\|C - A\| \approx 0.204 < \|B - A\| \approx 0.212 \tag{23}$$

and

$$D_{\mathrm{KL}}(B, A) \approx 0.0589 < D_{\mathrm{KL}}(C, A) \approx 0.069. \tag{24}$$

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 & Pettigrew, 2010, 206) and $E = W$ (all possible worlds are epistemically accessible),

$$\mathrm{GExp}_A(C) \approx 0.653 < \mathrm{GExp}_A(B) \approx 0.656. \tag{25}$$

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

One way to understand evidence is that it excludes a set of credences. Evidence which results in standard conditioning, for example, excludes all credences which assign a value less than 1 to the evidence E. Richard Jeffrey has remarked that more realistically evidence is not certain, so instead of $P'(E) = 1$ my evidence may be that $P'(E) \neq P(E)$. In this case, Jeffrey has proposed what we now call Jeffrey conditioning, for which there are dynamic coherence arguments (see Armendt (1980) and Goldstein (1983); and Skyrms, 1986), met with critical resistance in the literature (see Levi (1987), Christensen (1999), Talbott (1991), and Maher (1992); and Howson and Urbach, 2006). Jeffrey conditioning conserves ratios of probabilities that are unaffected by the evidence. In the Holmes example, b retains a 3:1 ratio between $P'(E_2)$ and $P'(E_3)$, while c does not; see Eq. 20.

Let me make more precise what I mean by recommendation. A scoring rule recommends an updated credence if the updated credence fulfills the following two

conditions: (i) it respects the evidence; and (ii) of all credences that respect the evidence, the updated credence is closest to the prior credence in terms of the scoring rule under consideration. In Imre Csiszár's terms, the updated credence is the projection of the prior credence onto the convex (or, less generally, affine) set of credences allowed by the intervening evidence. If the evidence is a convex set of permissible credences, then the recommendation is unique. This is why convexity is sometimes required for evidence; both standard conditioning type evidence and Jeffrey type evidence trivially fulfill this requirement as they are affine. Under this description of recommendation and belief update, the Log score fulfills RATIO CONSERVATION while the Brier score violates it.

I will make two claims here whose proof I omit for brevity (once the formal apparatus is estabished, these proofs are not difficult). (1) The Log Score does not uniquely fulfill RATIO CONSERVATION. There are many other Bregman divergences that fulfill it. (2) RATIO CONSERVATION is equivalent to CARDINALITY INDEPENDENCE plus permutation invariance. In the Holmes example, E_1 is the event that Mr. R is the culprit; E_2 is the event that Ms. S is the culprit; E_3 is the event that Ms. T is the culprit. If Holmes concludes by some evidence that $P'(E_1) = 1/2$, he needs to update $P'(E_2)$ and $P'(E_3)$ accordingly, for example by Jeffrey conditioning.

Now let us say that Holmes breaks down the outcome space differently. E_1 and E_2 remain as before, but E_3 is broken down into E_{3a} and E_{3b} by a conjunction with an unrelated event. For example, E_{3a} may be the event that Ms. T is the culprit and the course of the pound rose or remained equal relative to the price of gold in the last 24 hours. E_{3b} may be the event that Ms. T is the culprit and the course of the pound sank relative to the price of gold in the last 24 hours. Holmes' update upon evidence should be invariant to such an artificial inflation of the outcome space. Leszek Wroński calls this expectation CARDINALITY INDEPENDENCE (see Wroński, 2016). It is easy to show that RATIO CONSERVATION implies CARDINALITY INDEPENDENCE.

To show the converse we need another requirement: updates do not operate differently on permutations. If (a, b, c) is updated to (a', b', c'), then (b, a, c) is updated to (b', a', c'). Permutation invariance and CARDINALITY INDEPENDENCE imply RATIO CONSERVATION by a simple continuity argument.

The Brier score famously violates CARDINALITY INDEPENDENCE (see Levinstein, 2012) and therefore violates RATIO CONSERVATION. The Log score fulfills RATIO CONSERVATION (and therefore also CARDINALITY INDEPENDENCE), but the Log score is not a unique Bregman divergence to do so.

4 Symmetry

A scoring rule S is symmetric when it is associated with a divergence D such that

$$D_S(p\|q) = D_S(q\|p) \text{ for all } p \text{ and } q. \tag{26}$$

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). Against Pettigrew, I maintain 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. Against Selten, I maintain that 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. 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 Section 5.1).

An anonymous referee helpfully pointed out an ambiguity with respect to my use of the word "strength." In the context of this paper, I call a belief strong if its entropy is low, independent of its evidentiary support. Because the entropy function of a proper scoring rule is concave, weakness has a global maximum; and, more generally, entropy induces a structure on beliefs which makes a coordinate-free approach feasible.

5 Locality

5.1 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). Leonard J. Savage has shown that the only scoring rules fulfilling LOCALITY are the Log score and its not relevantly different close relatives.

Example 5.1 (Tokens) Before Casey draws one token 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

$$T(p_2, \ldots, p_{n-1}) = 1 - 2p_1 + \sum_{i=1}^{n-1} p_i^2 + \left(1 - \sum_{i=1}^{n-1} p_i\right)^2. \tag{27}$$

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.

5.2 Bronfman objection

Here is how LOCALITY may still work in favour of information theory against the geometry of reason. 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); for the original article see Bronfman (2009)).

Example 5.2 (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 does 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.

Let a uniqueness claim have dependent and independent reasons: choosing from a set of epistemic procedures, procedure X is unique for these reasons. The dependent reasons justify the uniqueness on account of the features that X exhibits. The independent reasons make no reference to these features, but provide a reason to have a unique successful candidate for winning the 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 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 presses 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.

6 Updating

6.1 Information theory and the geometry of reason

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

Evidence appears in the form of a constraint on acceptable probability distributions and the closest acceptable probability to the original (relatively prior)

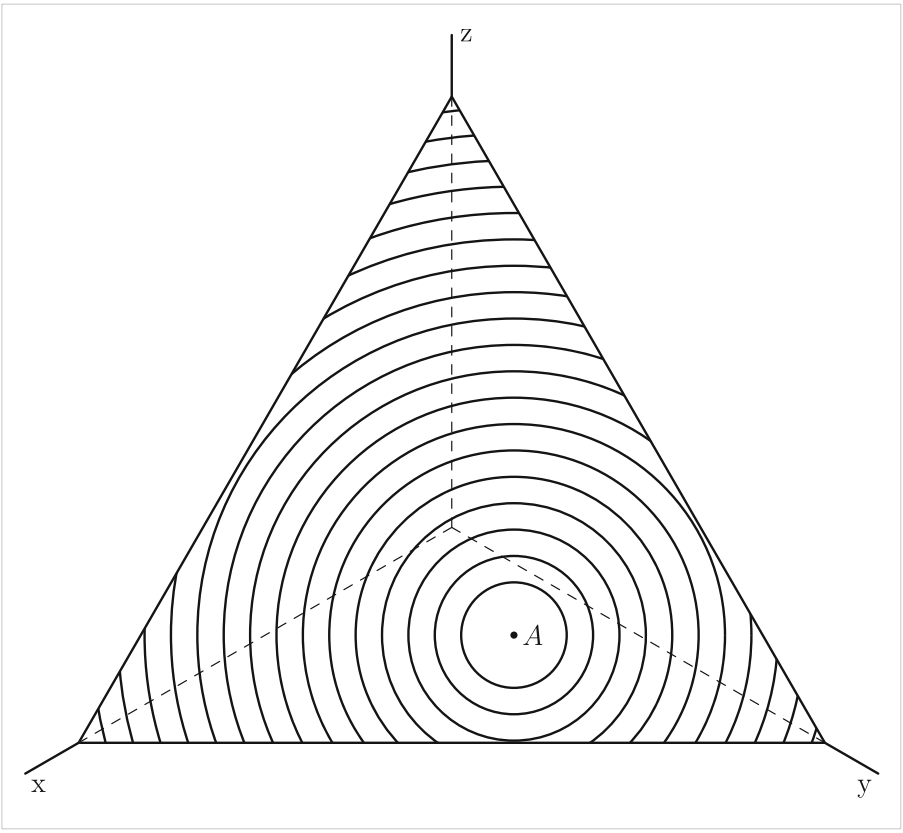

Fig. 1 The simplex $\mathbb{S}^2$ in three-dimensional space $\mathbb{R}^3$ with contour lines corresponding to the geometry of reason around point A in Eq. 20. Points on the same contour line are equidistant from A with respect to the Euclidean metric. The contour lines of the geometry of reason are insensitive to the boundaries of the simplex, while the contour lines of information theory reflect them (see Fig. 2). Note that this diagram and all the following diagrams are frontal views of the simplex

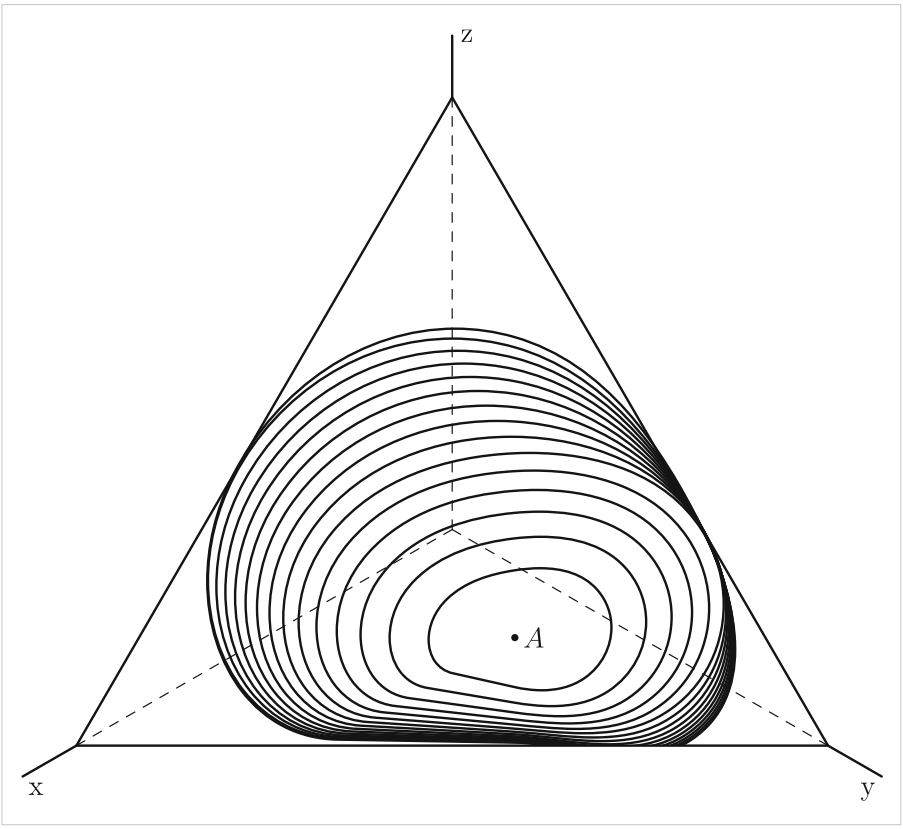

Fig. 2 The simplex $\mathbb{S}^2$ with contour lines corresponding to information theory around point A in Eq. 20. Points on the same contour line are equidistant from A with respect to the Kullback-Leibler divergence. Information theory respects epistemic intuitions we have about asymmetry: proximity to extreme beliefs with very high or very low probability influences the topology that is at the basis of updating. However, my overall argument is that representing information theory within coordinate geometry as in this diagram is the central error, not the preference of the Brier score over the Log score

probability distribution is chosen as its successor. As long as I insist on the coordinated approach, here is a list of reasonable expectations I may have toward this concept of quantitative difference $d(p, q)$ (we call it a distance function for the geometry of reason and a divergence for information theory).

- 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.
- LITTORAL COMPRESSION $d(p, q)$ has a tendency to measure distributions near the centre as being closer together than distributions near the extremes. This requirement is difficult to formalize without antecedent assumptions about distance or divergence, thus begging the question. One hope to make this more

precise is the theory of convex analysis (see (Rockafellar, 1997)), which naturally nudges us from plain to differential geometry.

- 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.

The Kullback-Leibler divergence of information theory fails all the expectations of this list except LITTORAL COMPRESSION. The Euclidean distance of the geometry of reason fulfills them except LITTORAL COMPRESSION. The shoe example and the bicycle example, as compelling as they are in their original context, may lack application to the kinematics of credences. I am, of course, not wedded to them, as they are deeply reflective of the coordinated approach, which I reject. They are not irrelevant in the sense that it takes effort to move from one belief state to another, as it takes effort to buy a shoe or ride a bicycle. I hope that to some readers they are helpful as illustrations of the absurdity while the coordinated approach is in play.

6.2 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 2.2) 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 (see Lukits, 2013). It is non-commutative and may provide the kind of asymmetry required to reflect epistemic asymmetry. However, it violates TRIANGULARITY and TRANSITIVITY OF ASYMMETRY. The task of this section is to show how serious these violations are. They dissipate once we no longer think of information theory operating on probability coordinates.

6.2.1 Triangularity

Let B be on the zero-sum line between A and C if and only if

$$d(A, C) = d(A, B) + d(B, C), \tag{28}$$

where d is the difference measure we are using, so $d(A, B) = \|B - A\|$ for the geometry of reason and $d(A, B) = D_{\mathrm{KL}}(B, A)$ for information geometry. For the geometry of reason (and Euclidean geometry), the zero-sum line between two probability distributions is just what we intuitively think of as a straight line: in Cartesian coordinates, B is on the zero-sum line strictly between A and C if and only if for some $\vartheta \in (0, 1)$, $b_i = \vartheta a_i + (1 - \vartheta)c_i$ and $i = 1, \ldots, n$.

What the zero-sum line looks like for information theory is illustrated in Fig. 3. 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 Eq. 28 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 $d(A, B) + d(B, C) < d(A, C)$.

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 concave 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:

$$d_H(x, y) \geq d_H(x, y') + d_H(y', y), \tag{29}$$

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 (see 2.2.5 in Grünwald & Dawid, 2004). When $\mathcal{C}$ is affine (such as a plane in $\mathbb{R}^3$), Eq. 29 turns from an inequality to an equation (replacing "$\geq$" by "$=$") for all Bregman divergences. For the squared Euclidean distance, Eq. 29 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 Eq. 20 violate TRIANGULARITY for D_{KL} because

$$0.067806 = D_{\mathrm{KL}}(A, B) + D_{\mathrm{KL}}(B, C) < D_{\mathrm{KL}}(A, C) = 0.071530. \tag{30}$$

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

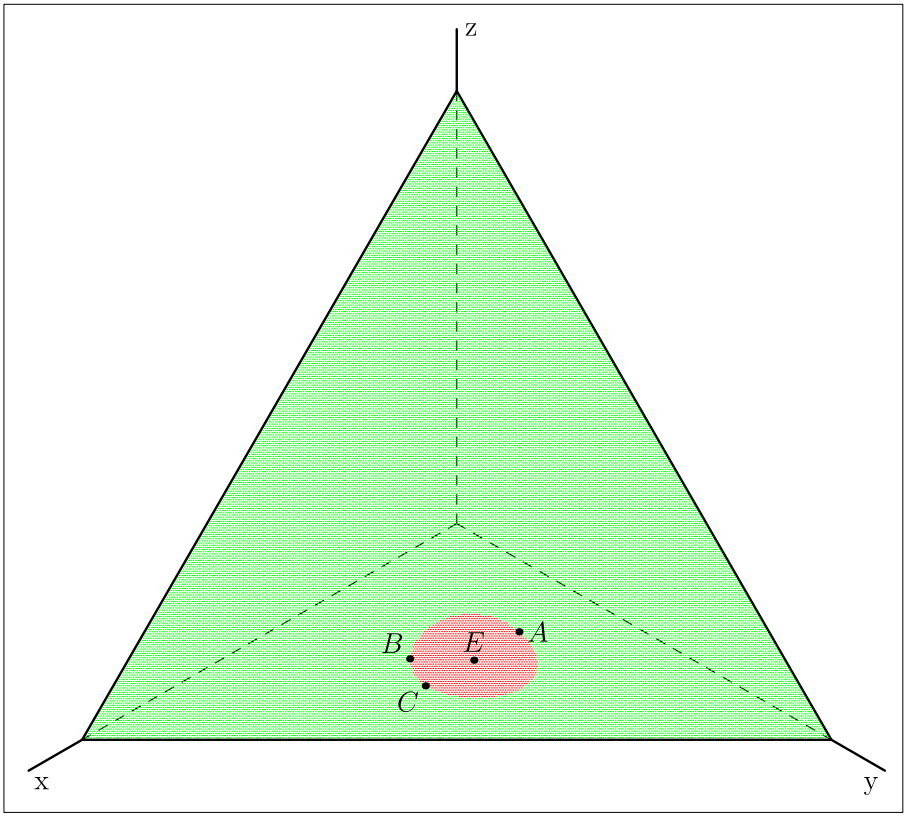

Fig. 3 The zero-sum line between A and C is the boundary line between the green area, where the triangle inequality holds, and the red area, where the triangle inequality is violated. The posterior probability distribution B recommended by Jeffrey conditioning always lies on the zero-sum line between the prior A and the Leitgeb Pettigrew posterior C (explained in Leitgeb & Pettigrew, 2010). 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, $e_k = \frac{c_k}{W\left(\frac{c_k}{\alpha_k}\exp(1+\lambda)\right)}$, 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 6.1 *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:*

$$D_{KL}(z, x) > D_{KL}(y, x) + D_{KL}(z, y)\,. \tag{31}$$

Proof It is straightforward to see that Eq. 31 is equivalent to

$$\sum_{i=1}^{n}(z_i - x_i)\log\frac{\vartheta x_i + (1-\vartheta)z_i}{x_i} > 0. \tag{32}$$

Expand the right hand side to

$$\sum_{i=1}^{n}\left(z_i+\frac{\vartheta}{1-\vartheta}x_i-\frac{\vartheta}{1-\vartheta}x_i-x_i\right)\log\frac{\frac{1}{1-\vartheta}\left(\vartheta x_i+(1-\vartheta)z_i\right)}{\frac{1}{1-\vartheta}x_i}>0. \tag{33}$$

Equation 33 is clearly equivalent to Eq. 32. It is also equivalent to

$$\sum_{i=1}^{n}\left(z_i+\frac{\vartheta}{1-\vartheta}x_i\right)\log\frac{z_i+\frac{\vartheta}{1-\vartheta}x_i}{\frac{1}{1-\vartheta}x_i}+\sum_{i=1}^{n}\frac{1}{1-\vartheta}x_i\log\frac{\frac{1}{1-\vartheta}x_i}{z_i+\frac{\vartheta}{1-\vartheta}x_i}>0, \tag{34}$$

which is true by Gibbs' inequality. □

Like Bregman divergences in general, the Kullback-Leibler divergence in particular violates TRIANGULARITY by design. Giving Proposition 6.1 a misguided and paradoxical reading from the intuitions of geometry, the more often you stop on the way, the faster you reach your destination.

6.2.2 Transitivity of asymmetry

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

Example 6.1 (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 and posterior probabilities are reversed, so $Y_2 = (0, 0.5, 0.5)$ and $Y_2' = (0.4, 0.3, 0.3)$.

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.

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, reflecting a coordinated approach. 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

$$\Delta_P(P') = D_{\mathrm{KL}}(P', P) - D_{\mathrm{KL}}(P, P'), \tag{35}$$

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

$$\Delta^{-1}(\mathbb{R}_{>0})\Delta^{-1}(\mathbb{R}_{<0})\Delta^{-1}(\{0\}). \tag{36}$$

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 (36) 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 Fig. 4 for illustration). A more coordinate-free approach using differential geometry may be helpful.

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)

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. I call this requirement TRANSITIVITY OF ASYMMETRY. The Kullback-Leibler divergence violates it.

Consider

$$P_1 = \left(\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\right) P_2 = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right) P_3 = \left(\frac{2}{5}, \frac{2}{5}, \frac{1}{5}\right). \tag{37}$$

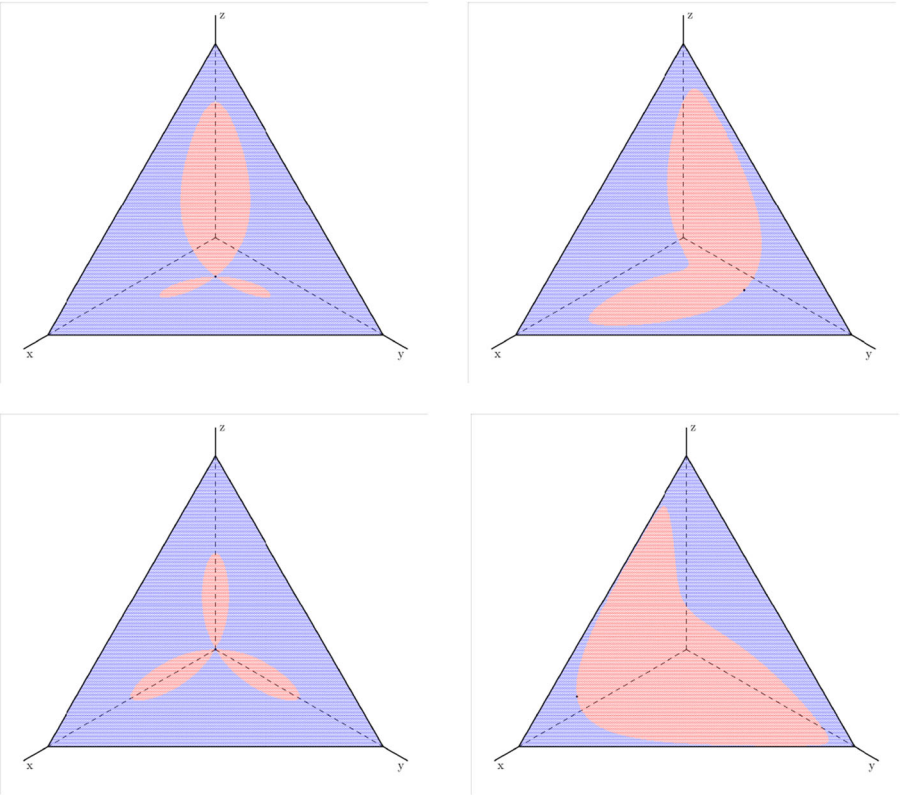

Fig. 4 The partition (36) based on different values for P. From top left to bottom right, $P = (0.4, 0.4, 0.2)$; $P = (0.242, 0.604, 0.154)$; $P = (1/3, 1/3, 1/3)$; $P = (0.741, 0.087, 0.172)$. Note that for the geometry of reason, the diagrams are trivial. The challenge for information theory is to explain the non-triviality of these diagrams epistemically without begging the question

(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 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 Vanommen et al. (2013, 193); Tobler's wind model in Tobler (1975); and the cyclist lovingly hand-sketched in Kopperman (1988, 91).

7 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 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 Section 2.1 I have summarized, however, that the geometry of reason fails several plausible requirements (INFORMATION, LOCALITY, LITTORAL COMPRESSION), 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 Section 5.2). 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. Information theory saddles us with a divergence function, the Kullback-Leibler divergence, which in the coordinated framework is excessively ill-behaved.

More promising is the use of differential geometry in the formal theory of partial beliefs. This paper primarily shows that entrenching partial beliefs as parametrized by probabilities $p_i, i = 1, \ldots, n$ is a dead-end road. 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

$$\left(\ln \frac{p_1}{p_n}, \ldots, \ln \frac{p_{n-1}}{p_n}\right)^\intercal. \tag{38}$$

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. What the abstractions make possible, perhaps by generalizing information theory, is a question for which I eagerly await answers.

Acknowledgments I want to thank Paul Bartha at the University of British Columbia and Franz Huber at the University of Toronto for their support as I worked on this paper.

Funding The author received financial support for the research, authorship, and publication of this article from the Social Sciences and Humanities Research Council, SSHRC award 756-2017-0286.

Declarations

Conflict of Interest The author declares no conflict of interest.

Ethical Approval The author declares that no ethical approval was needed for this paper and the study was conducted in accordance with the Helsinki Declaration.

Informed Consent The author declares that no informed consent was needed for this paper.

References

Armendt, B. (1980). Is there a dutch book argument for probability kinematics?. *Philosophy of Science*, *47*(4), 583–588.

Borges, J. (1962). *Ficciones*. New York, NY: Grove Press.

Bregman, L. (1967). 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), 200–217.

Bronfman, A. (2009). A gap in joyce's argument for probabilism. University of Michigan: unpublished manuscript.

Chino, N. (1978). A graphical technique for representing the asymmetric relationships between n objects. *Behaviormetrika*, *5*(5), 23–44.

Christensen, D. (1999). Measuring confirmation. *Journal of Philosophy*, *96*(9), 437–461.

Coombs, C. H. (1964). *A theory of data*. New York: Wiley.

Coxon, A. (1982). *The user's guide to multidimensional scaling*. Exeter: Heinemann Educational Books.

Csiszár, I., & Shields, P. C. (2004). *Information theory and statistics: A tutorial*. Hanover: Now Publishers.

Dawid, A. P., Lauritzen, S., & Parry, M. (2012). Proper local scoring rules on discrete sample spaces. *Annals of Statistics*, *40*(1), 593–608.

De Finetti, B. (2017). *Theory of probability*. Chichester: Wiley.

Gentleman, R., Ding, B., Dudoit, S., & Ibrahim, J. (2006). Distance measures in dna microarray data analysis. In R. Gentleman, V. Carey, W. Huber, R. Irizarry, & S. Dudoit (Eds.) *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer.

Goldstein, M. (1983). The prevision of a prevision. *Journal of the American Statistical Association*, *78*(384), 817–819.

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

Grünwald, P. D., & Dawid, A. P. (2004). Game theory, maximum entropy, minimum discrepancy and robust bayesian decision theory. *The Annals of Statistics*, *32*(4), 1367–1433.

Hájek, A. (2003). What conditional probability could not be. *Synthese*, *137*(3), 273–323.

Harshman, R., & Lundy, M. (1984). The parafac model for three-way factor analysis and multidimensional scaling. In H. G. Law (Ed.) *Research methods for multimode data analysis* (pp. 122–215). New York: Praeger.

Harshman, R. A., Green, P. E., Wind, Y., & Lundy, M.E. (1982). A model for the analysis of asymmetric data in marketing research. *Marketing Science*, *1*(2), 205–242.

Hendrickson, A., & Buehler, R. (1971). Proper scores for probability forecasters. *Annals of Mathematical Statistics*, *42*(6), 1916–1921.

Howson, C. (2008). De finetti, countable additivity, consistency and coherence. *British Journal for the Philosophy of Science*, *59*(1), 1–23.

Howson, C., & Urbach, P. (2006). *Scientific reasoning: The bayesian approach*, 3rd edn. Chicago: Open Court.

Joyce, J. (2015). The value of truth: A reply to howson. *Analysis*, *75*(3), 413–424.

Kopperman, R. (1988). All topologies come from generalized metrics. *American Mathematical Monthly*, *95*(2), 89–97.
Landes, J. (2015). Probabilism, entropies and strictly proper scoring rules. *International Journal of Approximate Reasoning*, *63*, 1–21.
Leitgeb, H., & Pettigrew, R. (2010). An objective justification of bayesianism i: Measuring inaccuracy. *Philosophy of Science*, *77*(2), 201–235.
Levi, I. (1987). The demons of decision. *The Monist*, *70*(2), 193–211.
Levinstein, B. A. (2012). Leitgeb and pettigrew on accuracy and updating. *Philosophy of Science*, *79*(3), 413–424.
Lukits, S. (2013). The principle of maximum entropy and a problem in probability kinematics. *Synthese*, 1–23.
MacKay, D. (2003). *Information theory, inference and learning algorithms*. Cambridge: Cambridge.
Maher, P. (1992). Diachronic rationality. *Philosophy of Science*, 120–141.
Marschak, J. (1959). *Remarks on the economics of information*. Technical Report, Cowles Foundation for Research in Economics, Yale University.
McCarthy, J. (1956). Measures of the value of information. *Proceedings of the National Academy of Sciences*, *42*(9), 654–655.
Miller, D. (1984). A geometry of logic. In H. Skala, S. Termini, & E. Trillas (Eds.) *Aspects of Vagueness* (pp. 91–104). Dordrecht: Reidel.
Mormann, T. (2005). Geometry of logic and truth approximation. *Poznan Studies in the Philosophy of the Sciences and the Humanities*, *83*(1), 431–454.
Paul, L. A. (2016). *Transformative experience*. Oxford: Oxford University.
Pemberton, M., & Rau, N. (2016). *Mathematics for economists: An introductory textbook*. Manchester: Manchester University Press.
Pettigrew, R. (2016). *Accuracy and the laws of credence*. Oxford: Oxford University.
Predd, J., Seiringer, R., Lieb, E., Osherson, D., Poor, H. V., & Kulkarni, S. (2009). Probabilistic coherence and proper scoring rules. *IEEE Transactions on Information Theory*, *55*(10), 4786–4792.
Rockafellar, R. (1997). *Convex analysis*. Princeton: Princeton University.
Savage, L. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, *66*(336), 783–801.
Selten, R. (1998). Axiomatic characterization of the quadratic scoring rule. *Experimental Economics*, *1*(1), 43–62.
Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(1), 379–423, 623–656.
Skyrms, B. (1986). Dynamic coherence. In *Advances in the Statistical Sciences: Foundations of Statistical Inference* (pp. 233–243). Springer.
Talbott, W. J. (1991). Two principles of bayesian epistemology. *Philosophical Studies*, *62*(2), 135–150.
Tobler, W. (1975). *Spatial interaction patterns*. Schloss Laxenburg: International Institute for Applied Systems Analysis.
van Ommen, S., Hendriks, P., Gilbers, D., van Heuven, V., & Gooskens, C. (2013). Is diachronic lenition a factor in the asymmetry in intelligibility between danish and swedish?. *Lingua*, *137*, 193–213.
Wroński, L. (2016). Belief update methods and rules: Some comparisons. *Ergo, 3*(11).