Anomalies of Information Geometry

Stefan Lukits

1 Introduction

2 Triangulating LP and Jeffrey Conditioning

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

$$d(A, C) = d(A, B) + d(B, C) \tag{1}$$

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 figure 1. The reason for the oddity is that the Kullback-Leibler divergence does not obey TRIANGULARITY, an issue that I will address in detail in subsection 4.1). Call B a zero-sum point between A and C if (1) 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 poins are the boundary points of the set where you can take a shortcut by making a detour, i.e. all points for which $d(A, B) + d(B, C) < d(A, C)$.

Remarkably, if A represents a relatively prior probability distribution and C the posterior probability distribution recommended by LP conditioning,

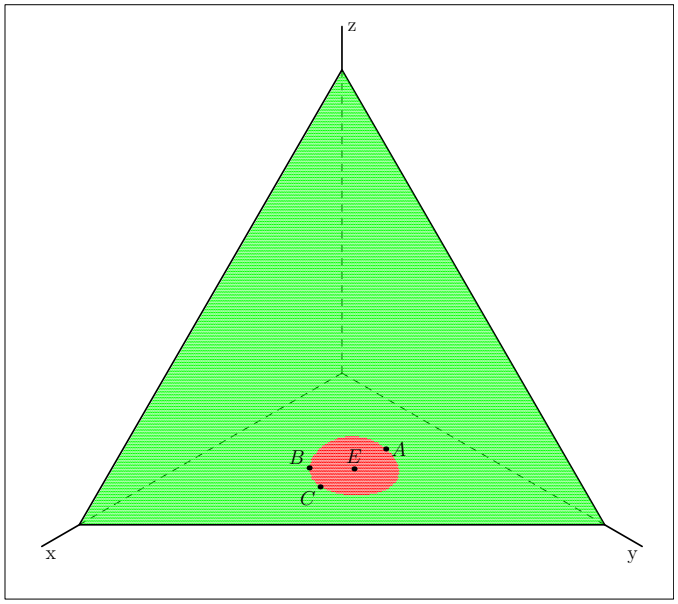

Figure 1: The zero-sum line between A and C is the boundary line between the green area, where the triangle inequality holds, and the red area, where the triangle inequality is violated. The posterior probability distribution B recommended by Jeffrey conditioning always lies on the zero-sum line between the prior A and the LP posterior C, as per equation (2). E is the point in the red area where the triangle inequality is most efficiently violated.

the posterior probability distribution recommended by Jeffrey conditioning is always a zero-sum point with respect to the Kullback-Leibler divergence:

$$D_{\mathrm{KL}}(C, A) = D_{\mathrm{KL}}(B, A) + D_{\mathrm{KL}}(C, B) \tag{2}$$

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.

To prove equation (2) in the case $n = 3$ (assuming that LP conditioning does not 'fall off the edge' as in case (b) in Leitgeb and Pettigrew, 2010, 253) note that all three points (prior, point recommended by Jeffrey conditioning, point recommended by LP conditioning) can be expressed using three variables:

$$\begin{aligned} A &= (1-\alpha, \beta, \alpha-\beta) \\ B &= \left(1-\gamma, \tfrac{\gamma\beta}{\alpha}, \tfrac{\gamma(\alpha-\beta)}{\alpha}\right) \\ C &= \left(1-\gamma, \beta+\tfrac{1}{2}(\gamma-\alpha), \alpha-\beta+\tfrac{1}{2}(\gamma-\alpha)\right) \end{aligned} \tag{3}$$

The rest is basic algebra using the definition of the Kullback-Leibler divergence. To prove the claim for arbitrary n one simply generalizes (3). It is a handy corollary of (2) that whenever (A, B) and (B, C) violate TRANSITIVITY OF ASYMMETRY then

$$D_{\mathrm{KL}}(A, C) > D_{\mathrm{KL}}(B, C) + D_{\mathrm{KL}}(A, B) \tag{4}$$

in violation of TRIANGULARITY. This way I will not have to go hunting for an example to demonstrate the violation of TRIANGULARITY. A, B, C of (5) fulfill all the conditions for (4) and therefore violate TRIANGULARITY.

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

It is an interesting question to wonder which point E violates the triangle inequality most efficiently so that

$$D_{\mathrm{KL}}(E, C) + D_{\mathrm{KL}}(A, E) \tag{6}$$

is minimal. Let $e = (e_1, \ldots, e_n)$ represent E in $\mathbb{S}^{n-1}$. Use the Lagrange Multiplier method to find the Lagrangian

$$\mathcal{L}(e, \lambda) = \sum_{i=1}^{n} e_i \log \frac{e_i}{a_i} + \sum_{i=1}^{n} c_i \log \frac{c_i}{e_i} + \lambda \left(\sum_{i=1}^{n} e_i - 1 \right) \tag{7}$$

The Lagrange Multiplier method gives us

$$\frac{\partial \mathcal{L}}{\partial e_k} = \log \frac{e_k}{a_k} + 1 - \frac{r}{q} + \lambda = 0 \text{ for each } k = 1, \ldots, n. \tag{8}$$

Manipulate this equation to yield

$$\frac{c_k}{e_k} \exp\left(\frac{c_k}{e_k}\right) = \frac{c_k}{a_k} \exp(1 + \lambda). \tag{9}$$

To solve (9), use the Lambert W function

$$e_k = \frac{c_k}{W\left(\frac{c_k}{a_k} \exp(1 + \lambda)\right)}. \tag{10}$$

Choose λ to fulfill the constraint $\sum e_i = 1$. The result for the discrete case accords with Ovidiu Calin and Constantin Udriste's result for the continuous case (see equation 4.7.9 in Calin and Udriste, 2014, 127). Numerically, for A and C as defined in equation (5),

$$E = (0.415, 0.462, 0.123). \tag{11}$$

This is subtly different from the midpoint $m_i = 0.5a_i + 0.5c_i$ (if we were minimizing $D_{\mathrm{KL}}(A, E) + D_{\mathrm{KL}}(C, E)$, the solution would be the midpoint). I do not know whether A, E, C are collinear (see figure 1 for illustration).

3 Confirmation

The geometry of reason thinks about the comparison of probability distributions in terms of distance. Information theory thinks about the comparison

along the lines of information loss when one distribution is used to encode a message rather than the other distribution. One way to test these approaches is to ask how well they align with a third approach to such a comparison: degree of confirmation. Our main concern is the horizon effect of the previous subsection. Which approaches to degree of confirmation theory reflect it, and how do these approaches correspond to the disagreements between information theory and the geometry of reason?

There is, of course, a relevant difference between the aims of the epistemic utility approach to updating and the aims of degree of confirmation theory. The former investigates norms to which a rational agent conforms in her pursuit of epistemic utility. The latter seeks to establish qualitative and quantitative measures of impact that evidence has on a hypothesis. Both, however, (I will restrict my attention here to quantitative degree of confirmation theory) attend to the probability of an event, which degree of confirmation theory customarily calls h for hypothesis, before and after the rational agent processes another event, customarily called e for evidence, i.e. $x = P(h|k)$ and $y = P(h|e, k)$ (k is background information).

For perspectives on the link between confirmation and information see Shogenji, 2012, 37f; Crupi and Tentori, 2014; and Milne, 2014, section 4. Vincenzo Crupi and Katya Tentori suggest that there is a "parallelism between confirmation and information search [which] can serve as a valuable heuristic for theoretical work" (Crupi and Tentori, 2014, 89).

In degree of confirmation theory, incremental confirmation is distinguished from absolute confirmation in the following sense. Let h be the presence of a very rare disease and e a test result such that $y >> x$ but $y < 1 - y$. Then, absolutely speaking, e disconfirms h (for Rudolf Carnap, absolute confirmation involves sending y above a threshold r which must be greater than or equal to 0.5). Absolute confirmation is not the subject of this section. I will exclusively discuss incremental confirmation (also called relevance confirmation, just as absolute confirmation is sometimes called firmness confirmation) where $y > x$ implies (incremental) confirmation, $y < x$ implies (incremental) disconfirmation, and $y = x$ implies the lack of both. The difference is illustrated in figure 2.

All proposed measures of quantitative, incremental degree of confirmation considered here are a function of x and y. Dependence of incremental confirmation on only x and y is not trivial, as $P(e|k)$ and $P(e|h, k)$ cannot be expressed using only x and y (for a case why dependence should be on only x

and y see Atkinson, 2012, 50, with an irrelevant conjunction argument; and Milne, 2014, 254, with a continuity argument). David Christensen's measure $P(h|e,k) - P(h|\neg e,k)$ (see Christensen, 1999, 449) and Robert Nozick's $P(e|h,k) - P(e|\neg h,k)$ (see Nozick, 1981, 252) are not only dependent on x and y, but also on $P(e|k)$, which makes them vulnerable to Atkinson's and Milne's worries just cited.

Consider the following six contenders for a quantitative, incremental degree of confirmation function, dependent on only x and y. They are based on, in a brief slogan, (i) difference of conditional probabilities, (ii) ratio of conditional probabilities, (iii) difference of odds, (iv) ratio of likelihoods, (v) Gaifman's treatment of Hempel's raven paradox, and (vi) conservation of contrapositivity and commutativity. Logarithms throughout this paper are assumed to be the natural logarithm in order to facilitate easy differentiation, although generally a particular choice of base (greater than one) does not make a relevant difference.

$$
\begin{array}{lrcl}
\text{(i)} & M_P(x,y) & = & y - x \\
\text{(ii)} & R_P(x,y) & = & \log \dfrac{y}{x} \\
\text{(iii)} & J_P(x,y) & = & \dfrac{y}{1-y} - \dfrac{x}{1-x} \\
\text{(iv)} & L_P(x,y) & = & \log \dfrac{y(1-x)}{x(1-y)} \\
\text{(v)} & G_P(x,y) & = & \log \dfrac{1-x}{1-y} \\
\text{(vi)} & Z_P(x,y) & = & \begin{cases} \dfrac{y-x}{1-x} & \text{if } y \geq x \\ \dfrac{y-x}{x} & \text{if } y < x \end{cases}
\end{array}
\tag{12}
$$

M_P is defended by Carnap, 1962; Earman, 1992; Rosenkrantz, 1994. R_P is defended by Keynes, 1921; Milne, 1996; Shogenji, 2012. J_P is defended by Festa, 1999. L_P is defended by Good, 1950; Good, 1983, chapter 14; Fitelson, 2006; Zalabardo, 2009. G_P is defended by Gaifman, 1979, 120, without the logarithm (I added it to make G_P more comparable to the other functions).

Z_P is defended by Crupi et al., 2007. For more literature supporting the various measures consult footnote 1 in Fitelson, 2001, S124; and an older survey of options in Kyburg, 1983.

To compare how these degree of confirmation measures align with the concept of difference between probability distributions for the purpose of updating it is best to look at derivatives as they reflect the rate of change from the middle to the extremes. This is how we capture the horizon effect requirement for two dimensions. One important difference between degree of confirmation theory and updating is that the former is concerned with a hypothesis and its negation whereas the latter considers all sorts of domains for the probability distribution (in this paper, I have restricted myself to a finite outcome space). As far as the analogy between degree of confirmation theory on the one hand and updating on the other hand is concerned, I only need to look at the two-dimensional case.

To discriminate between candidates (i)–(vi), I am setting up three criteria (complementing many others in the literature). Let $D(x, y)$ be the generic expression for the degree of confirmation function. Call this **List C**.

- ADDITIVITY A theory can be confirmed piecemeal. Whether the evidence is split up into two or more components or left in one piece is irrelevant to the amount of confirmation it confers. Formally, $D(x, z) = D(x, y) + D(y, z)$. Note that this is not the usual triangle inequality because I am in two dimensions.

- SKEW-ANTISYMMETRY It does not matter whether h or $\neg h$ is in view. Confirmation and disconfirmation are commensurable. Formally, $D(x, y) = -D(1 - x, 1 - y)$. A surprising number of candidates fail this requirement, and the requirement is not common in the literature (see, however, the second clause in Milne's fourth desideratum in 1996, 21). In defence of this requirement consider example 1 below. $d_1 > d_2$ may have a negative impact on the latter scientist's grant application, even though the inequality may solely be due to a failure to fulfill skew-antisymmetry.

- CONFIRMATION HORIZON An account of degree of confirmation must exhibit the horizon effect as in **List A** and **List B**, except more simply in two dimensions. Formally, the functions $\partial D^+_\varepsilon/\partial x$ must be strictly positive and the functions $\partial D^-_\varepsilon/\partial x$ must be strictly negative for all $\varepsilon \in (-1/2, 1/2)$. These functions are defined in (13) and (14).

Example 1: Grant Adjudication I. Two scientists compete for grant money. Professor X presents an experiment conferring degree of confirmation d_1 on a hypothesis, if successful; Professor Y presents an experiment conferring degree of disconfirmation $-d_2$ on the negation of the same hypothesis, if unsuccessful. (For the relevance of quantitative confirmation measures to the evaluation of scientific projects see Salmon, 1975, 11.)

The functions for the horizon effect are defined as follows. Let $\varepsilon \in (-1/2, 1/2)$ be fixed. Recall that $D(x, y)$ is the generic expression for a confirmation function measuring the degree of confirmation that a posterior $y = P(h|e, k)$ bestows on a hypothesis for which the prior is $x = P(h|k)$. ε is the difference $y - x$. For $\varepsilon > 0$,

$$\begin{array}{lllll} D^-_\varepsilon : (0, \frac{1}{2} - \varepsilon) \to \mathbb{R} & D^-_\varepsilon(x) & = & |D(x, x+\varepsilon)| \\ D^+_\varepsilon : (\frac{1}{2}, 1 - \varepsilon) \to \mathbb{R} & D^+_\varepsilon(x) & = & |D(x, x+\varepsilon)| \end{array} \tag{13}$$

For $\varepsilon < 0$,

$$\begin{array}{lllll} D^-_\varepsilon : (-\varepsilon, \frac{1}{2}) \to \mathbb{R} & D^-_\varepsilon(x) & = & |D(x, x+\varepsilon)| \\ D^+_\varepsilon : (\frac{1}{2} - \varepsilon, 1) \to \mathbb{R} & D^+_\varepsilon(x) & = & |D(x, x+\varepsilon)| \end{array} \tag{14}$$

The rate of change for the different quantitative measures of degree of confirmation can be observed in figure 2. The pass and fail verdicts in the table below are evident from figure 2 and the table of derivatives provided in TBA. Only J_P, L_P and Z_P fulfill the horizon requirement.

Candidate	*Triangularity*	*Skew-Antisymmetry*	*Confirmation Horizon*
M_P	pass	pass	fail
R_P	pass	fail	fail
J_P	pass	fail	pass
L_P	pass	pass	pass
G_P	pass	fail	fail
Z_P	fail	pass	pass

The table makes clear that only L_P passes all three tests. I am not making a strong independent case for L_P here, especially against M_P, which is the most likely hero of the geometry of reason. This has been done elsewhere (for example in Schlesinger, 1995, where L_P and M_P are compared to each other

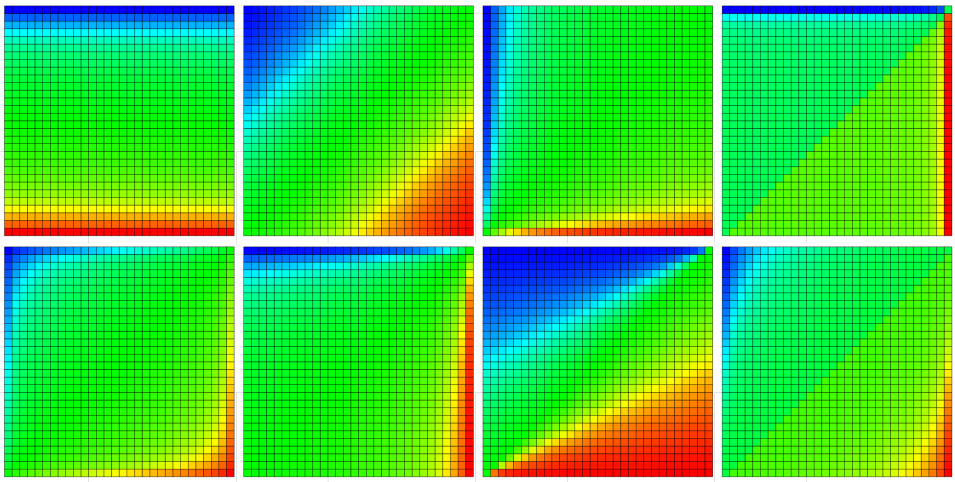

Figure 2: Illustration for the six degree of confirmation candidates plus Carnap's firmness confirmation and the Kullback-Leibler divergence. The top row, from left to right, illustrates FMRJ, the bottom row LGZI. 'F' stands for Carnap's firmness confirmation measure $F_P(x, y) = \log(y/(1-y))$. 'M' stands for candidate (i), $M_P(x, y)$ in (12), the other letters correspond to the other candidates (ii)-(v). 'I' stands for the Kullback-Leibler divergence multiplied by the sign function of $y - x$ to mimic a quantitative measure of confirmation. For all the squares, the colour reflects the degree of confirmation with x on the x-axis and y on the y-axis, all between 0 and 1. The origin of the square's coordinate system is in the bottom left corner. Blue signifies strong confirmation, red signifies strong disconfirmation, and green signifies the scale between them. Perfect green is $x = y$. G_P looks like it might pass the horizon requirement, but the derivative reveals that it fails CONFIRMATION HORIZON.

in their performance given some intuitive examples; Elliott Sober presents the counterargument in Sober, 1994). The argumentative force of this subsection appeals to those who are already sympathetic to L_P. Adherents of M_P will hopefully find other items on **List A** persuasive and reject the geometry of reason, in which case they may come back to this subsection and re-evaluate their commitment to M_P.

> **Example 2: Grant Adjudication II.** Two scientists compete for grant money. Professor X presents an experiment that will increase the probability of a hypothesis from 98% to 99%, if successful. Professor Y presents an experiment that will increase the probability of a hypothesis from 1% to 2%, if successful.

All else being equal, Professor Y should receive the grant money. If her

experiment is more successful, it will arguably make more of a difference. This example illustrates that the analogy between degree of confirmation and updating remains tenuous, since for degree of confirmation theory the consensus on intuitions is far inferior to the updating case. If, for example, the confirmation function is anti-symmetric and $D(x, y) - D(y, x)$ is zero (for M_P and L_P, for example), then together with skew-antisymmetry this means that degree of confirmation is equal for Professor X and Professor Y. Despite its three passes in the above table, L_P fails here.

Based on Roberto Festa's J_P, Professor X's prospective degree of confirmation is 5000 times larger than Professor Y's, but Festa in particular insists "that there is no universally applicable P-incremental c-measure, and that the appropriateness of a P-incremental c-measure is highly context-dependent" (Festa, 1999, 67). R_P and Z_P appear to be sensitive to example 2. The Kullback-Leibler divergence gives us the right result as well, where the degree of confirmation for going from 1% to 2% is $3.91 \cdot 10^{-3}$ compared to $3.12 \cdot 10^{-3}$ for going from 98% to 99%, but the Kullback-Leibler divergence is not a serious degree of confirmation candidate. It fulfills SKEW-ANTISYMMETRY and CONFIRMATION HORIZON, but not ADDITIVITY (see subsection 4.1).

Intuitions easily diverge here. Christensen may be correct when he says, "perhaps the controversy between difference and ratio-based positive relevance models of quantitative confirmation reflects a natural indeterminateness in the basic notion of 'how much' one thing supports another" (Christensen, 1999, 460). Pluralists allow therefore for "distinct, complementary notions of evidential support" (Hájek and Joyce, 2008, 123). I am sympathetic towards this indeterminateness in degree of confirmation theory, but not when it comes to updating (see the full employment theorem in Lukits, 2013, 1413).

This subsection assumes that despite these problems with the strength of the analogy, degree of confirmation and updating are sufficiently similar to be helpful in associating options with each other and letting the arguments in each other's favour and disfavour cross-pollinate. As an aside, Christensen's S-support given by evidence E is stable over Jeffrey conditioning on $[E, \neg E]$; LP-conditioning is not (see Christensen, 1999, 451). This may serve as another argument from degree of confirmation theory in favour of information theory (which supports Jeffrey conditioning) against the geometry of reason (which supports LP conditioning).

Consider the following two conditions on a difference measure D on a simplex $\mathbb{S}^{n-1} \subset \mathbb{R}^n$, which is assumed to be a smooth function from $\mathbb{S}^{n-1} \times \mathbb{S}^{n-1} \to \mathbb{R}$.

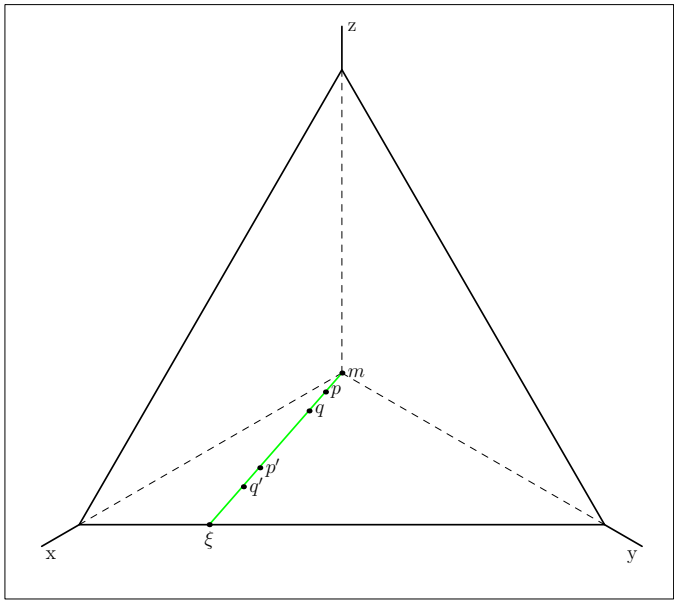

Figure 3: An illustrations of conditions (i)–(iii) for COLLINEAR HORIZON in **List B**. p, p' and q, q' must be equidistant and collinear with m and ξ. If q, q' is more peripheral than p, p', then COLLINEAR HORIZON requires that $|d(p, p')| < |d(q, q')|$.

(h1) 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 condition see figure 3.

(h2) Let $\mu \in (-1, 1)$ be fixed and D_μ defined as in (16). Then $dD_\mu/dx > 0$, where dD_μ/dx is the total derivative as x moves towards $\xi(x)$, the unique boundary point which is collinear with x and m.

To define D_μ, the hardest part is to specify the domain. Let this domain $V(\mu) \subseteq \mathbb{S}^{n-1}$ be defined as

$$V(\mu) = \begin{cases} \{x \in \mathbb{S}^{n-1} | x_i < (1-\mu)\xi_i(x) + \mu m_i, i = 1, \ldots, n\} & \text{for } \mu > 0 \\ \{x \in \mathbb{S}^{n-1} | x_i > (1+\mu)m_i - \mu\xi_i(x), i = 1, \ldots, n\} & \text{for } \mu < 0. \end{cases} \tag{15}$$

Then $D_\mu : V(\mu) \to \mathbb{R}_0^+$ is defined as

$$D_\mu(x) = |D(x, y(x))| \tag{16}$$

where $y_i(x) = x_i + \mu(\xi_i(x) - m_i)$. Remember that $\xi(x)$ is the unique boundary point which is collinear with x and m. Now for the proof that (h1) and (h2) are equivalent.

First assume (h1) and the negation of (h2). Since D is smooth, there must be a $\bar{\mu}$ and two points x' and x'' collinear with m and a boundary point $\bar{\xi}$ such that $D_{\bar{\mu}}(x') \geq D_{\bar{\mu}}(x'')$ even though $\|\bar{\xi} - x''\| < \|\bar{\xi} - x'\|$. If this were not the case, D_μ would be strictly increasing running towards the boundary points for all μ and its total derivative would be strictly positive so that (h2) follows. Now consider the four points x', x'', y', y'' where $y_i' = x_i' + \mu(\bar{\xi}_i - m_i)$ and $y_i'' = x_i'' + \mu(\bar{\xi}_i - m_i)$ for $i = 1, \ldots, n$. Without loss of generality, assume $\bar{\mu} > 0$. Then x', x'', y', y'' fulfill the conditions in (h1) and $D_{\bar{\mu}}(x') < D_{\bar{\mu}}(x'')$, in contradiction to the aforesaid.

Then assume (h2). Let x', x'', y', y'' be four points as in (h1). Consider $\mu = \|\xi - m\|/\|x'' - x'\|$. Then $D_\mu(x') = |D(x', x'')|$ and $D_\mu(y') = |D(y', y'')|$. (h2) tells us that along a path from m to ξ, D_μ is strictly increasing, so $D_\mu(x') = |D(x', x'')| < |D(y', y'')| = D_\mu(y')$. QED.

Note that the Euclidean distance function violates both (h1) and (h2) in all dimensions. The Kullback-Leibler divergence fulfills them if $n = 2$ but violates them if $n > 2$. For $n = 2$, this is easily checked by considering the derivative of the Kullback-Leibler divergence for two dimensions (use D_ε defined in (13) and (14) for the two-dimensional case instead of D_μ for the arbitrary-dimensional case). A counterexample to fulfillment of (h1) and (h2) for $n = 3$ is given in (24).

Now that I have shown that (h1) and (h2) is equivalent, it is easy to show that J_P, L_P and Z_P fulfill the horizon requirement while M_P, R_P and G_P violate it. The following table of derivatives will do (note that $\varepsilon = y - x$ is

fixed while x varies and that these derivatives have to be considered with the absolute value for the various degree of confirmation functions in mind). Column 1 is the name of the candidate confirmation function; column 2 is the function with x and $\varepsilon = y - x$ as arguments; column 3 is the derivative $\partial D(x,\varepsilon)/\partial x$ for $|D(x,\varepsilon)| = D(x,\varepsilon)$.

$M_P(x,y)$	ε	0
$R_P(x,y)$	$\log \frac{x+\varepsilon}{x}$	$-\frac{\varepsilon}{(x+\varepsilon)x}$
$J_P(x,y)$	$\frac{x+\varepsilon}{1-x-\varepsilon} - \frac{x}{1-x}$	$\frac{1}{(1-x-\varepsilon)^2} - \frac{1}{(1-x)^2}$
$L_P(x,y)$	$\log \frac{(x+\varepsilon)(1-x)}{x(1-x-\varepsilon)}$	(17)
$G_P(x,y)$	$\log \frac{1-x}{1-x-\varepsilon}$	$\frac{h}{(1-x)(1-x-\varepsilon}$
$Z_P(x,y)$	$\frac{\varepsilon}{1-x}$ or $\frac{\varepsilon}{x}$	$\frac{\varepsilon}{(1-x)^2}$ or $-\frac{\varepsilon}{x^2}$

with

$$-\frac{((x+\varepsilon)(1-x)-(1-x-\varepsilon)x)(1-2x-\varepsilon)}{(x+\varepsilon)(1-x-\varepsilon)(1-x)x}. \tag{17}$$

4 Expectations for Information Theory

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

The Kullback-Leibler divergence, which is the most promising concept of difference for probability distributions in information theory and the one which gives us Bayesian standard conditioning as well as Jeffrey condition-

ing, is non-commutative and may provide the kind of asymmetry required to reflect epistemic asymmetry. However, it also violates TRIANGULARITY, COLLINEAR HORIZON, and TRANSITIVITY OF ASYMMETRY. The task of this section is to show how serious these violations are.

4.1 Triangularity

As mentioned at the end of section 2, the three points A, B, C in (5) violate TRIANGULARITY as in (4):

$$D_{\mathrm{KL}}(A, C) > D_{\mathrm{KL}}(B, C) + D_{\mathrm{KL}}(A, B). \tag{18}$$

This is counterintuitive on a number of levels, some of which I have already hinted at in illustration: taking a shortcut while making a detour; buying a pair of shoes for more money than buying the shoes individually.

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

$$\delta(x, z) > \delta\left(x, y\right) + \delta\left(y, z\right). \tag{19}$$

I will prove this in a moment, but here is a disturbing consequence: think about an ever more finely grained sequence of partitions y^j, $j \in \mathbb{N}$, of the line segment from x to z with y^{jk} as dividing points. I will spare myself defining these partitions, but note that any dividing point $y^{j_0 k}$ will also be a dividing point in the more finely grained partitions y^{jk} with $j \geq j_0$. Then define the sequence

$$T_j = \sum_k \delta\left(y^{jk}, y^{j(k+1)}\right) \tag{20}$$

such that the sum has as many summands as there are dividing points for j, plus one (for example, two dividing points divide the line segment into

three possibly unequal thirds). If δ were the Euclidean distance norm, T_j would be constant and would equal $\|z - x\|$. Zeno's arrow moves happily along from x to z, no matter how many stops it makes on the way. Not so for information theory and the Kullback-Leibler divergence. According to (19), any stop along the way reduces the sum of divergences.

T_j is a strictly decreasing sequence (does it go to zero? – I do not know, but if yes, it would add to the poignancy of this violation). The more stops you make along the way, the closer you bring together x and z.

For the proof of (19), it is straightforward to see that (19) is equivalent to

$$\sum_{i=1}^{n}(z_i - x_i)\log\frac{\vartheta x_i + (1-\vartheta)z_i}{x_i} > 0. \tag{21}$$

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

$$\sum_{i=1}^{n}\left(z_i + \frac{\vartheta}{1-\vartheta}x_i - \frac{\vartheta}{1-\vartheta}x_i - x_i\right)\log\frac{\frac{1}{1-\vartheta}\left(\vartheta x_i + (1-\vartheta)z_i\right)}{\frac{1}{1-\vartheta}x_i} > 0. \tag{22}$$

(22) is clearly equivalent to (21). It is also equivalent to

$$\sum_{i=1}^{n}\left(z_i + \frac{\vartheta}{1-\vartheta}x_i\right)\log\frac{z_i + \frac{\vartheta}{1-\vartheta}x_i}{\frac{1}{1-\vartheta}x_i} + \sum_{i=1}^{n}\frac{1}{1-\vartheta}x_i\log\frac{\frac{1}{1-\vartheta}x_i}{z_i + \frac{\vartheta}{1-\vartheta}x_i} > 0, \tag{23}$$

which is true by Gibbs' inequality.

4.2 Collinear Horizon

There are two intuitions at work that need to be balanced: on the one hand, the geometry of reason is characterized by simplicity, and the lack of curvature near extreme probabilities may be a price worth paying; on the other hand, simple examples such as example 3 make a persuasive case for curvature.

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

Information theory is characterized by a very complicated 'semi-quasimetric' (the attribute 'quasi' is due to its non-commutativity, the attribute 'semi' to its violation of the triangle inequality). One of its initial appeals is that it performs well with respect to the horizon requirement near the boundary of the simplex, which is also the location of Schlesinger's examples. It is not trivial, however, to articulate what the horizon requirement really demands.

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

$$p = \left(\frac{1}{5}, \frac{2}{5}, \frac{2}{5}\right) \qquad p' = q = \left(\frac{1}{4}, \frac{3}{8}, \frac{3}{8}\right) \qquad q' = \left(\frac{3}{10}, \frac{7}{20}, \frac{7}{20}\right) \tag{24}$$

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

$$D_{\mathrm{KL}}(B, A) = 7.3820 \cdot 10^{-3} > 6.4015 \cdot 10^{-3} = D_{\mathrm{KL}}(C, B). \tag{25}$$

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

Figure 4 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 $x_i = 0.5(m_i + y_i)$.

$$\text{(i) } D_{\mathrm{KL}}(Y, M) > D_{\mathrm{KL}}(M, Y) \text{ and (ii) } D_{\mathrm{KL}}(X, M) > D_{\mathrm{KL}}(Y, X) \tag{26}$$

In terms of coordinates, the inequalities reduce to

$$\text{(i) } H(y) < \frac{1}{n}\sum(\log y_i) - \log\frac{1}{n^2} \text{ and} \tag{27}$$

$$\text{(ii) } H(y) > \log\frac{4}{n} - \sum\left[\left(\frac{3}{2}y_i + \frac{1}{2n}\right)\log\left(y_i + \frac{1}{n}\right)\right]. \tag{28}$$

(i) is simply the case described in the next subsection for asymmetry and illustrated on the bottom left of figure 5. (ii) tells us how far from the midpoint I can go with a scenario where $p = m, p' = q$ while violating COLLINEAR HORIZON. Clearly, as illustrated in figure 4, there is a relationship between asymmetry and COLLINEAR HORIZON.

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

4.3 Transitivity of Asymmetry

Recall Joyce's two axioms Weak Convexity and Symmetry. The geometry of reason (certainly in its Euclidean form) mandates Weak Convexity because the bisector of an isosceles triangle is always shorter than the isosceles sides. Weak Convexity also holds for information theory. Symmetry, however, fails for information theory. Fortunately, although I do not pursue this any further here, information theory arrives at many of Joyce's results even without the violated axiom.

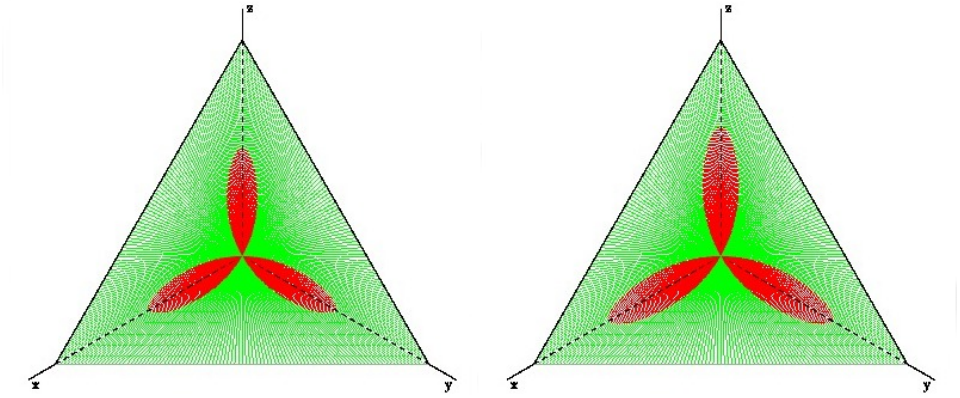

Figure 4: These two diagrams illustrate inequalities (27) and (28). The former displays all points in red which violate COLLINEAR HORIZON, measured from the centre. The latter displays points in different colours whose orientation of asymmetry differs, measured from the centre. The two red sets are not the same, but there appears to be a relationship, one that ultimately I suspect to be due to the more basic property of asymmetry.

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

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

not with what Alan Hájek calls the ratio analysis of conditional probability, see Hájek, 2003). The geometry of reason may want to solve this problem by signing on to a version of regularity, but then it violates REGULARITY. Happy kids, clean house, sanity: the hapless homemaker must pick two. The third remains elusive. Continuity, a consistent view of regularity, and symmetry: the hapless geometer of reason cannot have it all.

Now turn to the woes of the information theorist. Given the asymmetric similarity measure of probability distributions that information theory requires (the Kullback-Leibler divergence), a prior probability distribution P may be closer to a posterior probability distribution Q than Q is to P if their roles (prior-posterior) are reversed. That is just what we would expect. The problem is that there is another posterior probability distribution R where the situation is just the opposite: prior P is further away from posterior R than prior R is from posterior P. And whether a probability distribution different from P is of the Q-type or of the R-type escapes any epistemic intuition.

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

In other words, if

$$\Delta_P(P') = D_{\mathrm{KL}}(P', P) - D_{\mathrm{KL}}(P, P'), \tag{29}$$

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

$$\Delta^{-1}(\mathbb{R}_{>0}) \qquad \Delta^{-1}(\mathbb{R}_{<0}) \qquad \Delta^{-1}(\{0\}) \tag{30}$$

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 (30) is non-trivial—some probability distributions are of the Q-type, some are of the R-type, and it is difficult to think of an epistemic distinction between them that does not already presuppose information theory. See figure 5 for graphical illustration of this point.

On any account of well-behaved and ill-behaved asymmetries, the Kullback-Leibler divergence is ill-behaved. Of the four axioms as listed by Ralph Kopperman for a distance measure d (see Kopperman, 1988, 89), the Kullback-Leibler divergence violates both symmetry and triangularity, making it a 'semi-quasimetric':

(m1) $d(x, x) = 0$

(m2) $d(x, z) \leq d(x, y) + d(y, z)$ (triangularity)

(m3) $d(x, y) = d(y, x)$ (symmetry)

(m4) $d(x, y) = 0$ implies $x = y$ (separation)

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

$$P_1 = \left(\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\right) \qquad P_2 = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right) \qquad P_3 = \left(\frac{2}{5}, \frac{2}{5}, \frac{1}{5}\right) \quad (31)$$

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 (31) 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, or at least classification (it would help, for example, to know that all reasonable non-commutative difference measures used for updating are ill-behaved). Kopperman's objective is primarily to rescue continuity, uniform continuity, Cauchy sequences, and limits for topologies induced by difference measures which violate triangularity, symmetry, and/or separation. Kopperman does not touch axiom (m1), while in the psychological literature (see especially Tversky, 1977) self-similarity is an important topic. This is why an initially promising approach to asymmetric modeling in Hilbert spaces by Chino (see Chino, 1978; Chino, 1990; Chino and Shiraiwa, 1993; and Saburi and Chino, 2008) will not help us to distinguish well-behaved and ill-behaved asymmetries between probability distributions.

The failure of Chino's modeling approach to make useful distinctions among asymmetric distance measures between probability distributions leads us to the more complex theory of information geometry and differentiable manifolds. Both the results of Shun-ichi Amari (see Amari, 1985; and Amari and Nagaoka, 2000) and Nikolai Chentsov (see Chentsov, 1982) serve to highlight the special properties of the Kullback-Leibler divergence, not without elevating the discussion to a level of mathematical sophistication, however, where it is difficult to retain the appeal to epistemic intuitions. Information geometry considers probability distributions as differentiable manifolds equipped with a Riemannian metric. This metric, however, is Fisher's information metric, not the Kullback-Leibler divergence, and it is defined on the tangent space of the simplex representing finite-dimensional probability distributions. There is a sense in which the Fisher information metric is the derivative of the Kullback-Leibler divergence, and so the connection to epistemic intuitions can be re-established.

For a future research project, it would be lovely 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 mea-

sure on which to base inference and updating. Chentsov's theory of monotone invariance and Amari's theory of α-connections are potential candidates to provide such results as well as an epistemic justification for information theory.

5 Joyce's Result Still Stands

Using information theory instead of the geometry of reason, Joyce's result still stands, vindicating probabilism on epistemic merits rather than prudential ones: partial beliefs which violate probabilism are dominated by partial beliefs which obey it, no matter what the facts are.

Joyce's axioms, however, will need to be reformulated to accommodate asymmetry. This appendix shows that the axiom Weak Convexity still holds in information geometry. Consider three points $Q, R, S \in \mathbb{S}^{n-1}$ (replace $\mathbb{S}^{n-1}$ by the n-dimensional space of non-negative real numbers, if you do not want to assume probabilism) for which

$$D_{\mathrm{KL}}(Q, R) = D_{\mathrm{KL}}(Q, S). \tag{32}$$

I will show something slightly stronger than Weak Convexity: Joyce's inequality is not only true for the midpoint between R and S but for all points $\vartheta R + (1 - \vartheta)S$, as long as $0 \leq \vartheta \leq 1$. The inequality aimed for is

$$D_{\mathrm{KL}}(Q, \vartheta R + (1 - \vartheta)S) \leq D_{\mathrm{KL}}(Q, R) = D_{\mathrm{KL}}(Q, S). \tag{33}$$

To show that it holds I need the log-sum inequality, which is a result of Jensen's inequality (for a proof of the log-sum inequality see Theorem 2.7.1 in Cover and Thomas, 2006, 31). For non-negative numbers $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$,

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} a_i \right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}. \tag{34}$$

(33) follows from (34) via

$$D_{\mathrm{KL}}(Q,R) = \vartheta D_{\mathrm{KL}}(Q,R) + (1-\vartheta) D_{\mathrm{KL}}(Q,S) =$$
$$\sum_{i=1}^{n} \left(\vartheta q_i \log \frac{\vartheta q_i}{\vartheta r_i} + (1-\vartheta) q_i \log \frac{(1-\vartheta) q_i}{(1-\vartheta) s_i} \right) \geq$$
$$\sum_{i=1}^{n} q_i \log \frac{q_i}{\vartheta r_i + (1-\vartheta) s_i} = D_{\mathrm{KL}}(Q, \vartheta R + (1-\vartheta) S). \tag{35}$$

I owe some thanks to physicist friend Thomas Buchberger for help with this proof. Interested readers can find a more general claim in Csiszár's Lemma 4.1 (see Csiszár and Shields, 2004, 448), which accommodates convexity of the Kullback-Leibler divergence as a special case.

6 Asymmetry in Two Dimensions

This appendix contains a proof that the threefold partition (30) of $\mathbb{S}^1$ is well-behaved, in contrast to the threefold partition of $\mathbb{S}^2$ as illustrated by figure 5. For the two-dimensional case, i.e. considering $p, q \in \mathbb{S}^1$ with $0 < p, q < 1, p + p' = 1$ and $q + q' = 1$,

$$\begin{array}{lllllll} \Delta_q(p) & > & 0 & \text{for} & |p-p'| & > & |q-q'| \\ \Delta_q(p) & = & 0 & \text{for} & |p-p'| & = & |q-q'| \\ \Delta_q(p) & < & 0 & \text{for} & |p-p'| & < & |q-q'| \end{array} \tag{36}$$

where $\Delta_q(p) = D_{\mathrm{KL}}(q,p) - D_{\mathrm{KL}}(p,q)$ and $D_{\mathrm{KL}}(p,q) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}$. Part of information theory's ill behaviour outlined in section 4.3 is that in the higher-dimensional case the partition does not follow the simple rule that higher entropy of P compared to Q implies that $\Delta_Q(P) > 0$ (Δ here defined as in (29)). In the two-dimensional case, however, this simple rule applies.

That a comparison in entropy $H(p) = -p \log p - (1-p) \log(1-p)$ between $H(p)$ and $H(q)$ corresponds to a comparison of $|p-p'|$ and $|q-q'|$ is trivial. The proof for (36) is straightforward given the following non-trivial lemma establishing a very tight inequality. Given that $p + p' = 1$ and $q + q' = 1$ and $p, q, p', q' > 0$ it is true that

$$\text{If } \log(p/q) > \log(q'/p') \text{ then } (p+q)\log(p/q) > (p'+q')\log(q'/p') \quad (37)$$

Let $x = p/q$ and $y = q'/p'$. I know that $x > y$ since $\log x > \log y$. Now I want to show $(p+q)\log x > (p'+q')\log y$. Note that $p = xq, q' = p'y, p+q = q(x+1)$, and $p'+q' = p'(y+1)$. Therefore,

$$q = \frac{1-y}{1-xy} \quad (38)$$

and

$$p' = \frac{1-x}{1-xy}. \quad (39)$$

What I want to show is that $x > y$ implies

$$\frac{1-y}{1-xy}(x+1)\log x > \frac{1-x}{1-xy}\log y. \quad (40)$$

Note that $f(x) = (1-x)^{-1}(x+1)\log x$ is increasing on $(0,1)$ and decreasing on $(1,\infty)$, and consider the following two cases:

(i) When $x < 1, y < 1$, (40) follows from the fact that f is increasing on $(0,1)$.

(ii) When $x > 1, y > 1$, (40) follows from the fact that f is decreasing on $(1,\infty)$.

Mixed cases such as $x > 1, y < 1$ do not occur, as for example $x > 1$ implies $y > 1$.

7 The Hermitian Form Model

Asymmetric MDS is a promising approach to classify asymmetries in terms of their behaviour. This subsection demonstrates that Chino's asymmetric

MDS, both spatial and non-spatial, fails to give us explanations for information theory's violation of TRANSITIVITY OF ASYMMETRY. I am choosing Chino's approach because it is the most general and most promising of all the different asymmetric MDS models (see, for example, Chino and Shiraiwa, 1993, where Chino manages to subsume many of the other approaches into his own).

Multi-dimensional scaling (MDS) visualizes similarity of individuals in datasets. Various techniques are used in information visualization, in particular to display the information contained in a proximity matrix. When the proximity matrix is asymmetrical, we speak of asymmetric MDS. These techniques can be spatial (see for example Chino, 1978), where the proximity relationships are visualized in two-dimensional or higher-dimensional space; or non-spatial (see for example Chino and Shiraiwa, 1993), where the proximity relationships are used to identify data sets with abstract spaces (in Chino's case, finite-dimensional complex Hilbert spaces) and metrics defined on them.

The spatial approach in two dimensions fails right away for information theory because it cannot visualize transitivity violations. The hope for other types of asymmetric MDS is that it would be able to distinguish between well-behaved and ill-behaved asymmetries and either exclude or identify better-behaved candidates than the Kullback-Leibler divergence for measuring the distance between probability distributions. I will use Chino's most sophisticated non-spatial account to show that asymmetric MDS cannot solve this problem. For other asymmetric MDS note that with the Hermitian Form Model Chino seeks to integrate and generalize over all the other accounts.

Assume a finity proximity matrix. I will work with two examples here to avoid the detailed and abstract account provided by Chino. The first example is

$$D = \begin{bmatrix} 0 & 2 & 3 \\ 3 & 0 & 1 \\ -1 & 2 & 0 \end{bmatrix} \tag{41}$$

and allows for easy calculations. The second example corresponds to (31), the example for transitivity violation where

$$\hat{D} = \begin{bmatrix} 0.0000 & 0.0566 & 0.0487 \\ 0.0589 & 0.0000 & 0.0499 \\ 0.0437 & 0.0541 & 0.0000 \end{bmatrix}, \tag{42}$$

and the elements of the matrix $\hat{d}_{jk} = D_{\mathrm{KL}}(P_j, P_k)$. Note that the diagonal elements are all zero, as no updating is necessary to keep the probability distribution constant.

Chino first defines a symmetric matrix S and a skew-symmetric matrix T corresponding to the proximity matrix such that $D = S + T$.

$$S = \frac{1}{2}(D + D') \text{ and } T = \frac{1}{2}(D - D'). \tag{43}$$

Note that D' is the transpose of D, S is a symmetric matrix, and T is a skew-symmetric matrix with $t_{jk} = -t_{kj}$. Next we define the Hermitian matrix

$$H = S + iT, \tag{44}$$

where i is the imaginary unit. H is a Hermitian matrix with $h_{jk} = \overline{h_{kj}}$. Hermitian matrices are the complex generalization of real symmetric matrices. They have special properties (see section 8.9 in Anton and Busby, 2003) which guarantee the existence of a unitary matrix U such that

$$H = U\Lambda U^*, \tag{45}$$

where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ with n the dimension of D and λ_k the k-th eigenvalue of H (theorem 8.9.8 in Anton and Busby, 2003). U is the matrix of eigenvectors with the k-th column being the k-th eigenvector. U^* is the conjugate transpose of U. Given example (41), the numbers look as follows:

$$H = \frac{1}{2}\begin{bmatrix} 0 & 5-i & 2+4i \\ 5+i & 0 & 3-i \\ 2-4i & 3+i & 0 \end{bmatrix} \tag{46}$$

and

$$U = \begin{bmatrix} 0.019 + 0.639i & -0.375 + 0.195i & 0.514 + 0.386i \\ 0.279 - 0.494i & -0.169 - 0.573i & 0.503 + 0.260i \\ -0.519 + 0.000i & 0.681 - 0.000i & 0.516 + 0.000i \end{bmatrix} \tag{47}$$

with $\Lambda = \mathrm{diag}(-3.78, 0.0715, 3.71)$. Λ is calculated using the characteristic polynomial $\lambda^3 - 14\lambda + 1$ of H. Notice that the characteristic polynomial is a depressed cubic (the second coefficient is zero), which facilitates computation and will in the end spell the failure of Chino's program for our purposes.

Given example (42), the numbers are

$$\hat{H} = \frac{1}{2} \begin{bmatrix} 0.0000 + 0.0000i & 0.0578 - 0.0011i & 0.046 + 0.003i \\ 0.0578 + 0.0011i & 0.0000 + 0.0000i & 0.052 - 0.002i \\ 0.0462 - 0.0025i & 0.0520 + 0.0021i & 0.000 + 0.000i \end{bmatrix} \tag{48}$$

and

$$\hat{U} = \begin{bmatrix} 0.351 - 0.467i & -0.543 + 0.170i & -0.578 - 0.006i \\ -0.604 + 0.457i & -0.201 - 0.169i & -0.598 + 0.002i \\ 0.290 - 0.000i & 0.779 + 0.000i & -0.555 + 0.000i \end{bmatrix} \tag{49}$$

with $\Lambda = \mathrm{diag}(-0.060, -0.045, 0.104)$.

Chino now elegantly shows how the decomposition of $H = U\Lambda U^*$ defines a seminorm on a vector space. Let $\phi(\zeta, \tau) = \zeta\Lambda\tau^*$. Then (i) $\phi(\zeta_1 + \zeta_2, \tau) = \phi(\zeta_1, \tau) + \phi(\zeta_2, \tau)$, (ii) $\phi(a\zeta, \tau) = a\phi(\zeta, \tau)$, and (iii) $\phi(\zeta, \tau) = \overline{\phi(\tau, \zeta)}$. These three conditions characterize an inner product on a finite-dimensional complex Hilbert space, but only if a fourth condition is met: positive (or negative) definiteness ($\phi(\zeta, \zeta) \geq 0$) for all ζ). One might hope that positive definiteness identifies the more well-behaved asymmetries by associating with them a finite-dimensional complex Hilbert space with the norm $\|\zeta\| = \sqrt{\phi(\zeta, \zeta)}$ defined on it (Chino himself speculatively mentioned this hope to me in personal communication).

The hope does not come to fruition. Without a non-trivial self-similarity relation, all seminorms defined as above are indefinite, and thus all cats grey in

the night. Not only are well-behaved and ill-behaved asymmetries indistinguishable by the light of this seminorm, even the seminorms for symmetry are indefinite. Not only does this not help our programme, it also puts a serious damper on Chino's, who never mentions the self-similarity requirement (which, given that we are dealing with a proximity matrix, is substantial).

Based on a theorem in linear algebra (see theorem 4.4.12 in Anton and Busby, 2003),

$$\sum_{j=1}^{n} \lambda_j = \mathrm{tr}(A) \tag{50}$$

whenever the λ_j are the eigenvalues of A. The reader can easily verify this theorem by noticing that the roots of the characteristic polynomial add up to the second coefficient (which is the trace of the original matrix). It is well-known that the eigenvalues of a Hermitian matrix are real-valued (theorem 8.9.4 in Anton and Busby, 2003), which is an important component for Chino to define the seminorm $\|\zeta\|$ with the help of ϕ. Unfortunately, using (50), the eigenvalues are not only real, but also add up to the trace of H, which is zero unless there is a non-trivial self-similarity relation.

Tversky entertains such self-similarity relations in psychology (see tbd), and Chino is primarily interested in applications in psychology. When the eigenvalues add up to zero, however, there will be positive and negative eigenvalues (unless the whole proximity matrix is the null-matrix), which renders the seminorm as defined by Chino indefinite. The Kullback-Leibler divergence is trivial with respect to self-similarity: $D_{\mathrm{KL}}(P, P) = 0$ for all P.

8 Conclusion

References

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

Amari, Shun-ichi, and Hiroshi Nagaoka. *Methods of Information Geometry*. Providence, RI: American Mathematical Society, 2000.

Anton, Howard, and Robert Busby. *Contemporary Linear Algebra*. New York, NY: Wiley, 2003.

Atkinson, David. "Confirmation and Justification." *Synthese* 184, 1: (2012) 49–61.

Calin, Ovidiu, and Constantin Udriste. *Geometric Modeling in Probability and Statistics*. Heidelberg, Germany: Springer, 2014.

Carnap, Rudolf. *Preface to the Second Edition of Logical Foundations of Probability*. University of Chicago, 1962.

Chentsov, Nikolai. *Statistical Decision Rules and Optimal Inference*. Providence, R.I: American Mathematical Society, 1982.

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

———. "A Generalized Inner Product Model for the Analysis of Asymmetry." *Behaviormetrika* 17, 27: (1990) 25–46.

Chino, Naohito, and Kenichi Shiraiwa. "Geometrical Structures of Some Non-Distance Models for Asymmetric MDS." *Behaviormetrika* 20, 1: (1993) 35–47.

Christensen, David. "Measuring Confirmation." *The Journal of Philosophy* 96, 9: (1999) 437–461.

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

Cover, T.M., and J.A. Thomas. *Elements of Information Theory*, volume 6. Hoboken, NJ: Wiley, 2006.

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

Crupi, Vincenzo, and Katya Tentori. "State of the Field: Measuring Information and Confirmation." *Studies in History and Philosophy of Science Part A* 47: (2014) 81–90.

Crupi, Vincenzo, Katya Tentori, and Michel Gonzalez. "On Bayesian Measures of Evidential Support: Theoretical and Empirical Issues." *Philosophy of Science* 74, 2: (2007) 229–252.

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

Earman, John. *Bayes or Bust?* Cambridge, MA: MIT, 1992.

Festa, R. "Bayesian Confirmation." In *Experience, Reality, and Scientific Explanation: Essays in Honor of Merrilee and Wesley Salmon*, edited by Merrilee H. Salmon, Maria Carla Galavotti, and Alessandro Pagnini, Dordrecht: Kluwer, 1999, 55–88.

Fitelson, Branden. "A Bayesian Account of Independent Evidence with Applications." *Philosophy of Science* 68, 3: (2001) 123–140.

———. "Logical Foundations of Evidential Support." *Philosophy of Science* 73, 5: (2006) 500–512.

Gaifman, Haim. "Subjective Probability, Natural Predicates and Hempel's Ravens." *Erkenntnis* 14, 2: (1979) 105–147.

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, Irving. *Good Thinking: The Foundations of Probability and Its Applications*. Minneapolis, MN: University of Minnesota, 1983.

Good, John Irving. *Probability and the Weighing of Evidence*. London, UK: Griffin, 1950.

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

Hájek, Alan, and James Joyce. "Confirmation." In *Routledge Companion to the Philosophy of Science*, edited by S. Psillos, and M. Curd, New York, NY: Routledge, 2008, 115–129.

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.

Keynes, John Maynard. *A Treatise on Probability*. London, UK: Macmillan, 1921.

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

Kyburg, Henry. "Recent Work in Inductive Logic." In *Recent Work in Philosophy*, edited by Tibor Machan, and Kenneth Lucey, Totowa, NJ: Rowman and Allanheld, 1983, 87–150.

Leitgeb, Hannes, and Richard Pettigrew. "An Objective Justification of Bayesianism II: The Consequences of Minimizing Inaccuracy." *Philosophy of Science* 77, 2: (2010) 236–272.

Lukits, Stefan. "The Principle of Maximum Entropy and a Problem in Probability Kinematics." *Synthese* 1–23.

Milne, Peter. "log[P(h/eb)/P(h/b)] Is the One True Measure of Confirmation." *Philosophy of Science* 63, 1: (1996) 21–26.

———. "Information, Confirmation, and Conditionals." *Journal of Applied Logic* 12, 3: (2014) 252–262.

Nozick, Robert. *Philosophical Explanations*. Cambridge, MA: Harvard University, 1981.

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.

Rosenkrantz, R.D. "Bayesian Confirmation: Paradise Regained." *British Journal for the Philosophy of Science* 45, 2: (1994) 467–476.

Saburi, S., and N. Chino. "A Maximum Likelihood Method for an Asymmetric MDS Model." *Computational Statistics & Data Analysis* 52, 10: (2008) 4673–4684.

Salmon, Wesley. "Confirmation and Relevance." In *Induction, Probability, and Confirmation*, edited by Grover Maxwell, and Robert Milford Anderson, Minneapolis, MI: University of Minnesota Press, 1975, 3–36.

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

Shogenji, Tomoji. "The Degree of Epistemic Justification and the Conjunction Fallacy." *Synthese* 184, 1: (2012) 29–48.

Sober, Elliott. "No Model, No Inference: A Bayesian Primer on the Grue Problem." In *Grue!: The New Riddle of Induction*, edited by Douglas Frank Stalker, Chicago: Open Court, 1994, 225–240.

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

Tversky, Amos. "Features of Similarity." *Psychological Review* 84, 4: (1977) 327–352.

Zalabardo, José. "An Argument for the Likelihood-Ratio Measure of Confirmation." *Analysis* 69, 4: (2009) 630–635.

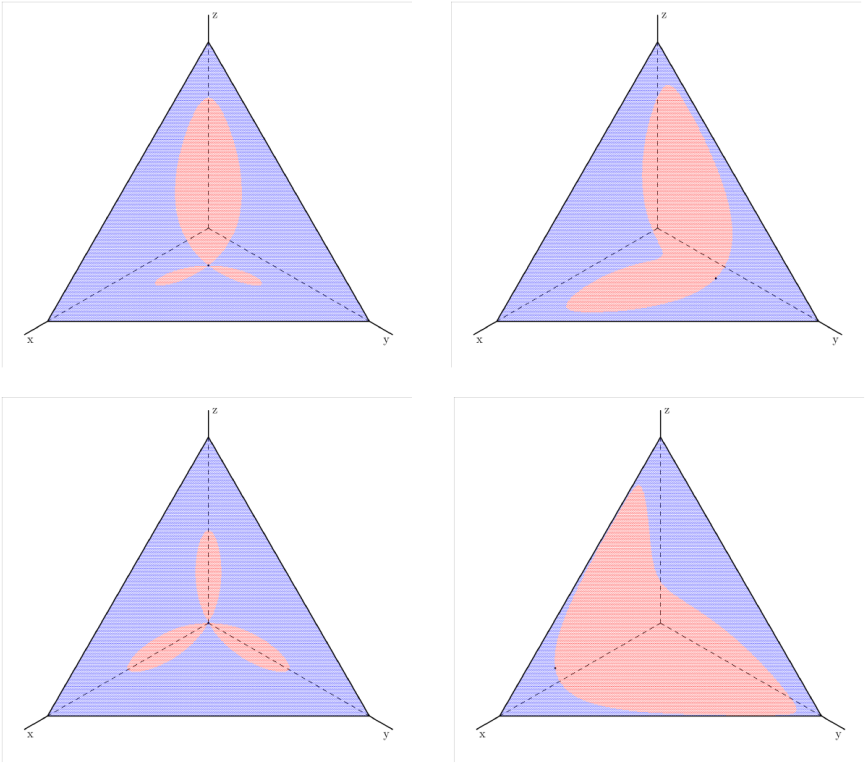

Figure 5: The partition (30) based on different values for P. From top left to bottom right, $P = (0.4, 0.4, 0.2); P = (0.242, 0.604, 0.154); P = (1/3, 1/3, 1/3); P = (0.741, 0.087, 0.172)$. Note that for the geometry of reason, the diagrams are trivial. The challenge for information theory is to explain the non-triviality of these diagrams epistemically without begging the question.