Failures of Information Geometry

John Skilling

Maximum Entropy Data Consultants Ltd, Kenmare, Ireland
skilling@eircom.net

Abstract. Information H is a unique relationship between probabilities, based on the property of *independence* which is central to scientific methodology. Information Geometry makes the tempting but fallacious assumption that a local metric (conventionally based on information) can be used to endow the space of probability distributions with a preferred global Riemannian metric.

No such global metric can conform to H, which is "from-to" asymmetric whereas geometrical length is by definition symmetric. Accordingly, *any* Riemannian metric will contradict the required structure of the very distributions which are supposedly being triangulated. Probabilities do not form a metric space.

We give counter-examples to alternative formulations of information, and to the use of information geometry.

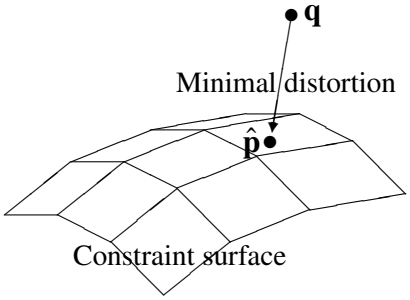
Keywords: Information geometry; metric space; probability distribution.
PACS: 02.50Cw;02.70Rr.

INFORMATION

The Bayesian sum and product rules allow us to do rational inference in accordance with a unique calculus [1, 2] which places probability on the unit simplex ($\sum_i p_i = 1$). The calculus is profitably extended by quantifying, as some function $H(\mathbf{p};\mathbf{q})$, the magnitude of change when a source distribution $\mathbf{q} = (q_1, q_2, \ldots)$ is updated to a destination distribution $\mathbf{p} = (p_1, p_2, \ldots)$.

$$\mathbf{p} \xleftarrow{\text{update}} \mathbf{q}$$

Usually, whatever constraints force change could be satisfied by a range of destinations. To remove this ambiguity, we ask that the chosen destination $\hat{\mathbf{p}}$ is a *minimal* distortion of the source $\mathbf{q}$.



$$\hat{\mathbf{p}} \xleftarrow[\text{minimise } H(\mathbf{p};\mathbf{q})]{\text{constraints}} \mathbf{q}$$

Other destinations might also satisfy the constraints, but would be "worse" in the sense of involving more distortion.

To uncover the form of H (if one exists), we use *independence*. A logician might quibble that there can never be true independence because everything's connected to

everything else. However, most connections are negligible, hence ignorable, otherwise we couldn't proceed at all. Problem A (winning my village lottery) doesn't noticeably influence problem B (getting a "six" next time I toss a die). Those two processes are deemed independent, and no practical consequence is expected if we choose to analyse them together.

$$\left.\begin{array}{l}\hat{\mathbf{p}}_A \longleftarrow \text{constraints } \mathbf{q}_A \text{ on } A \\ \hat{\mathbf{p}}_B \longleftarrow \text{constraints } \mathbf{q}_B \text{ on } B\end{array}\right\} \quad \equiv \quad \hat{\mathbf{p}}_A \times \hat{\mathbf{p}}_B \longleftarrow \text{constraints } \mathbf{q}_A \times \mathbf{q}_B$$

The unique "$p\log(p/q)$" information formula (generally attributed to Shannon [3] by physicists and to Kullback and Leibler [4] by statisticians) follows:

$$\left\| \quad H(\mathbf{p};\mathbf{q}) = \sum_i p_i \log \frac{p_i}{q_i} \quad \right\| \qquad \textbf{(information)} \tag{1}$$

We see that $H \geq 0$ with equality if and only if $\mathbf{p} = \mathbf{q}$, so it quantifies the distortion of $\mathbf{p}$ away from an arbitrary source distribution $\mathbf{q}$. This formula holds for arbitrary probabilities, and it satisfies independence. Hence the sought function H can exist, and it takes this uniquely defined form.

Minimising any other function leads to interference between independent applications, and that's unacceptable in a calculus of inference. Generalising the truth is a mistake which necessarily admits counter-examples.

Alternative proposals

Unfortunately, the definition of information remains questioned. Perhaps the term "entropy" (related to the negative of information) has caused confusion.

In physics, entropy quantifies the uncertainty about a system's state that remains after macroscopic constraints (on volume, temperature and so on) are applied. The combinatorics of a macroscopic system with independent components quickly lead to a "$-\sum p\log p$" entropy, and it's tempting to view this as a justification of that formula. Actually, it's no more than a sanity check, because any system with independence necessarily conforms. Conversely, systems lacking independence must and do have different formulas for their entropy. But that does not justify using different formulas for the information H from which those formulas ultimately derive.

The most popular alternative formula, invented without derivation, is

$$\text{⚠} \qquad H^{\dagger}_{\alpha}(\mathbf{p};\mathbf{q}) = \frac{1}{\alpha(1-\alpha)}\left(1 - \sum_i p_i^{\alpha} q_i^{1-\alpha}\right) \qquad \text{⚠} \tag{2}$$

as propounded by Rényi [5] and by Tsallis [6]. There are various special cases:

$\alpha = 2$	Least squares	$\frac{1}{2}\sum(p-q)^2/q$
$\alpha \to 1$	Information	$\sum p\log(p/q)$
$\alpha = \frac{1}{2}$	$\frac{1}{2}$(Hellinger distance)^2	$2\sum(\sqrt{p}-\sqrt{q})^2$
$\alpha \to 0$	Reverse information	$\sum q\log(q/p)$

(All these formulas have easy generalisations to non-normalised distributions.) We proceed to test the outcomes of minimising Rényi-Tsallis in various situations.

First counter-example

Consider the direct product of two probability distributions, $\mathbf{p}^{(1)} = (\frac{1}{10}, \frac{9}{10})$ and $\mathbf{p}^{(2)} = (\frac{1}{6}, \frac{5}{6})$. My chance of winning the village lottery is 1 in 10, and my chance of a "six" when I next throw a die is 1 in 6. Minimising the information (1) relative to uniform source $\mathbf{q}$ correctly produces the direct-product result $\mathbf{p}^{(1)} \times \mathbf{p}^{(2)} = (\frac{1}{10}, \frac{9}{10}) \times (\frac{1}{6}, \frac{5}{6})$.

Rényi-Tsallis does not. With $\alpha = 2$, which is least-squares, that result would involve a negative value if least-squares were taken seriously. In practice, positivity would supervene and force a hard zero.

$\frac{1}{10}$	$\frac{9}{10}$

$\frac{1}{6}$	
$\frac{5}{6}$	

$\frac{1}{60}$	$\frac{9}{60}$
$\frac{5}{60}$	$\frac{45}{60}$

What's expected

$\frac{-7}{60}$	$\frac{17}{60}$
$\frac{13}{60}$	$\frac{37}{60}$

or

0	$\frac{1}{6}$
$\frac{1}{10}$	$\frac{11}{15}$

What's delivered

The zero value indicates that winning the village lottery would prevent me throwing "six" with my next die — an implication that defies common sense.

Second counter-example

Consider the distribution of unit mass

$$M = \int_0^1 dx \int_0^1 dy\, p(x,y) = 1 \tag{3}$$

across the a-priori-uniform unit square $(0,1)\times(0,1)$. Known moments

$$\langle x \rangle = \int_0^1 dx \int_0^1 dy\, p(x,y)\, x = \frac{1}{6}, \qquad \langle y \rangle = \int_0^1 dx \int_0^1 dy\, p(x,y)\, y = \frac{1}{6} \tag{4}$$

constrain the centre of mass to $\langle (x,y) \rangle = (\frac{1}{6}, \frac{1}{6})$. This is in no way a difficult dataset.

Take $\alpha \to 0$. Minimising $H_0^\dagger$ under these constraints yields

$$\hat{p}(x,y) = 0.5379 \underbrace{\delta(x)\delta(y)}_{\text{mass}=1} + 0.4621 \underbrace{\frac{(\log 4)^{-1}}{x+y}}_{\text{mass}=1} \qquad \text{⚠} \tag{5}$$

with over half the mass concentrated into a delta-function singularity at the exact corner. This solution, inaccessible to any setting of Lagrange multipliers, would be rejected by

any thoughtful user, who would object to the coarse constraint producing an infinitely sharp result.

Systematic misbehaviour

Misbehaviour occurs whenever $\alpha \neq 1$. Minimising $H_\alpha^\dagger$ under integral constraints

$$\langle f_k \rangle = \int f_k(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} \tag{6}$$

yields

$$\hat{p}(\mathbf{x}) = \big(F(\mathbf{x})\big)^{1/(\alpha-1)} \tag{7}$$

where $F = \sum \lambda_k f_k$ is linear in the f's, with Lagrange multipliers λ as coefficients. For constraints so weak as to be ineffective, $F \approx 1$. For less weak constraints, F stays positive everywhere so that $\hat{p}$ is bounded. But, as the constraints require greater non-uniformity, the minimum value of F may shrink to zero.

When $\alpha > 1$, the consequence is that the density $\hat{p}$ becomes zero. As the constraints are strengthened even more, the Lagrange-multiplier solution (7) cannot respond without sending $\hat{p}$ negative. That being prohibited, a hard zero is imposed at the minimum of F. Each such zero, as it comes into play, removes the influence of $H_\alpha^\dagger$ until none remains. There is still a 1:1 correspondence between constraint values and the optimal $\hat{p}$, but duality with Lagrange multipliers λ fails because multipliers no longer characterise the result.

$$\text{constraints} \quad \overset{\text{1:1}}{\longleftrightarrow} \quad \hat{p} \quad \overset{\text{fail}}{\longleftrightarrow} \quad \lambda$$

When $\alpha < 1$, the consequence is that the density $\hat{p}$ becomes infinite. In a space of suitably high dimension, this can happen *without* the constraint values $\langle f_k \rangle$ becoming singular, as volumetric factors stabilise the infinite density by giving it finite mass. As the constraints are strengthened even more, the Lagrange-multiplier solution (7) cannot respond without sending $\hat{p}$ negative. Again, duality fails. Instead, a delta-function singularity is imposed at the minimum of F, which absorbs any further added mass.

Conclusion regarding the information formula

Scientific methodology requires results to be tested, and if (as here) a proposal fails simple tests, it cannot be recommended for complicated work. Danger lies not in simple problems where an immediate absurdity will guard the user against accepting error, but in more complicated situations where the consequences may be disguised and insidious.

Generalising the truth by ignoring relevant criteria (here, independence) damages it, and necessarily yields unacceptable results. This presages similar difficulties that arise when information is misinterpreted as geometry.

For inference, the only acceptable value for the Rényi-Tsallis parameter is $\alpha = 1$, which is the correct information (1). That negates the generalisation to $\alpha \neq 1$ which underlies Amari's "α-divergences" [7] in information geometry.

GEOMETRY

Being a smooth function, H necessarily has a symmetric second derivative

$$\frac{\partial^2 H}{\partial p_i \partial p_j} = \frac{\partial^2 H}{\partial p_j \partial p_i} = \frac{\delta_{ij}}{p_i} \tag{8}$$

which is widely used as a Riemannian metric g_{ij} in an identification usually attributed to Fisher [8] and Rao [9]. There, the length element $d\ell$ is defined by

$$(d\ell)^2 = \sum_{ij} g_{ij} dp_i dp_j = \sum_{ij} \frac{\partial^2 H}{\partial p_i \partial p_j} dp_i dp_j = \sum_i \frac{(dp_i)^2}{p_i} \tag{9}$$

Geodesic curves and lengths, and densities, are then constructed in the standard way, with microscopic local triangulation promoted to the macroscopic level.

Paths, lengths, density

The geodesic path from $\mathbf{q}$ to $\mathbf{p}$, linearly parameterised by θ and confined to the unit simplex, is

$$x_i = \left(\frac{\sin(\theta\gamma)}{\sin\gamma} \sqrt{p_i} + \frac{\sin((1-\theta)\gamma)}{\sin\gamma} \sqrt{q_i} \right)^2 \tag{10}$$

where $\gamma = \arccos(\sum_i \sqrt{p_i q_i})$. Its length

$$\ell(\mathbf{p}, \mathbf{q}) = 2\gamma \tag{11}$$

is basically Rényi-Tsallis with $\alpha = \frac{1}{2}$, and is somewhat greater than the Hellinger distance $4\sin(\gamma/2)$ which would be accessible if paths could leave the simplex. Meanwhile, the density over the unit simplex is

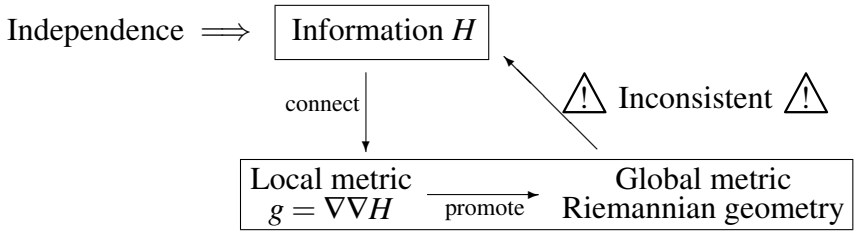
$$\rho(\mathbf{p}) \propto \delta\Big(\sum_i p_i - 1 \Big) \prod_i p_i^{-1/2} \tag{12}$$

Fundamental inconsistency

The connective H is "from-to" directed and not symmetric: $H(\mathbf{p};\mathbf{q}) \neq H(\mathbf{q};\mathbf{p})$. Its uniqueness implies that no acceptable symmetric connective exists. Geometric distance can be artificially endowed on the space, but any such distance is symmetric by construction, $\ell(\mathbf{p};\mathbf{q}) = \ell(\mathbf{q};\mathbf{p})$. So, any definition of geometric distance is necessarily incompatible with the independence that is at the heart of probabilistic practice.

$\|$ *Probabilities do not form a metric space.* $\|$

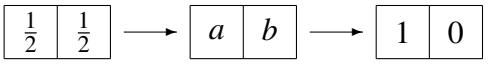
More precisely, imposition of a distance is incompatible with independence, and it's simply not possible to do science if irrelevant independent unknowns can't be discarded without changing the results.



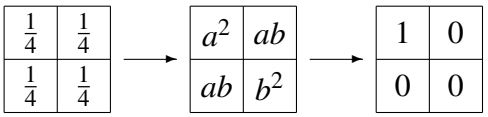
Awkward consequences must follow, and they do, as will be seen.

Geodesic paths

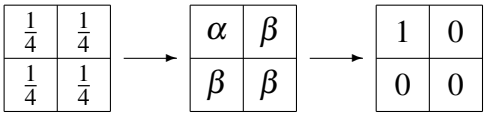
Consider a simple 2-cell probability problem, in which a path starts at $\mathbf{q} = (\frac{1}{2}, \frac{1}{2})$ and ends at $\mathbf{p} = (1,0)$. Normalisation only allows one degree of freedom, so there's only one track, (a,b) with $a+b=1$, which the geodesic must follow.



Now take the direct product of this problem with a second problem, which happens to be the same, so the product path starts at $(\frac{1}{2}, \frac{1}{2}) \times (\frac{1}{2}, \frac{1}{2})$ and ends at $(1,0) \times (1,0)$. Here is the independence path:



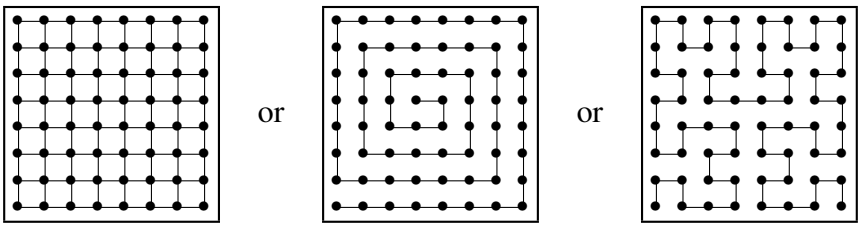
But the geodesic path, with three of the four cells starting the same and ending the same, is shown below with only two distinct values ($\alpha + 3\beta = 1$) instead of three.



Geometry does not distinguish between the three "β" cells. This elementary example demonstrates that geodesic paths do not conform to the independence that the informed user of probability might expect.

Start and finish points $\mathbf{q}$ and $\mathbf{p}$ do not in themselves define a unique path between them. In fact, the basic Bayesian task of learning about the contents of a domain does not even require a dimension, let alone a geometry. Thus the answers are the same whether a unit square is decomposed in two dimensions, or as a one-dimensional spiral, or some

quite different pattern. The choice is arbitrary, and usually made for computational convenience rather than reference to a supposedly pre-eminent geometry.



Conclusion regarding geodesic paths

Geometry is not fundamental to Bayesian analysis or computation, and in fact the freedom to discard topology and geometry is used to advantage in the general-purpose nested-sampling algorithm [10].

Geodesic length

Whether or not the path is confined to the simplex, the distance between two probability distributions is determined by $\sum_i \sqrt{p_i q_i}$, which is basically the Rényi-Tsallis formula (2) with $\alpha = \frac{1}{2}$. Accordingly, misbehaviour is expected.

Take the geometrically-defined closest probability distribution $\hat{\mathbf{p}}$ to uniform $\mathbf{q}$, subject to expectation

$$\int p(\mathbf{x}) E(\mathbf{x})\, d\mathbf{x} = \langle E \rangle \tag{13}$$

where

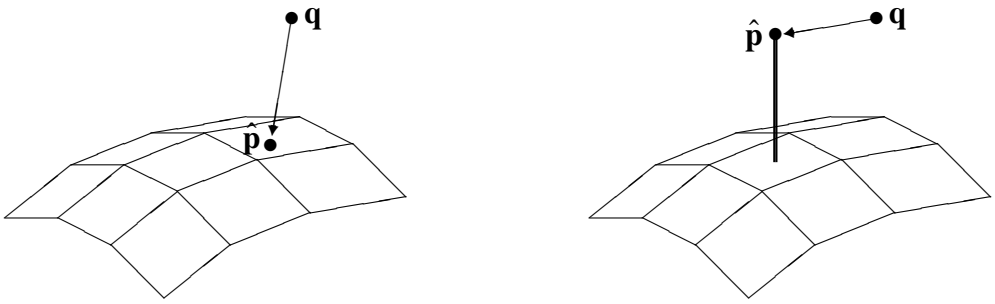
$$E(\mathbf{x}) = \sin^2(\pi x_1) + \sin^2(\pi x_2) + \sin^2(\pi x_3) + \sin^2(\pi x_4) + \sin^2(\pi x_5) + \sin^2(\pi x_6) \tag{14}$$

over the 6-dimensional unit cube $(-\frac{1}{2}, \frac{1}{2})^6$. The constraint value is $\langle E \rangle = 1$, implying a degree of central condensation towards the minimum $E_{\min} = 0$. This could represent a particle in a 6-dimensional periodic unit cell, or perhaps two particles in a 3-dimensional box, or 6 particles in a 1-dimensional box.

Minimising the information (1) would yield the smooth exponential form $\hat{p}(\mathbf{x}) \propto \exp(-3.650\, E(\mathbf{x}))$ familiar to physicists as the maximum-entropy distribution. Geometrically, though, the closest-to-uniform distribution is

$$\hat{p}(\mathbf{x}) = 0.5481 \underbrace{\delta(\mathbf{x})}_{\text{mass}=1} + 0.4519 \underbrace{\frac{5.944}{E(\mathbf{x})^2}}_{\text{mass}=1} \qquad \text{⚠} \tag{15}$$

with over half of the probability mass confined to a delta-function spike at the exact centre.



What's expected *What's delivered*

Conclusion regarding geodesic lengths

Informed users would not accept this infinite-resolution implication being drawn from a coarse constraint. Was one of the particles in a 3-dimensional box really definitively located at the exact centre? Did the average-energy constraint really support an infinite compression, quantified by $H(\mathbf{p};\mathbf{q}) = \infty$ bits of information?

Geometric density

Next, we test the suggestion that the $\sqrt{\det g}$ geometrical density could be a plausible assignment of belief (prior probability $\Pr(\mathbf{p})$), in a development taken forward by Amari [7] and followers.

First counter-example: Three proportions

The geometric prior for proportions $\mathbf{p} = (p_1, p_2, p_3)$ that add to 1 is

$$\Pr(\mathbf{p}) = \frac{1}{2\pi} \frac{\delta(p_1+p_2+p_3-1)}{\sqrt{p_1 p_2 p_3}} \tag{16}$$

Accurate observation yields a likelihood

$$\Pr(\text{data} \mid \mathbf{p}) = \delta(p_1 - p_3) \tag{17}$$

(If this delta function gives concern, use $\mathbf{1}(|p_1 - p_3| < \varepsilon)$ before taking $\varepsilon \to 0$.)

Perhaps masses 1 and 3 happened to balance. Perhaps the average number of spots $\langle n_s \rangle = p_1 + 2p_2 + 3p_3$ converged on 2 after many throws of a 3-die. There could be many applications: here we are concerned with the joint distribution

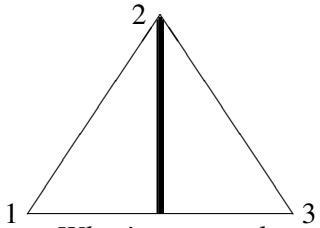
$$\Pr(\text{data}, \mathbf{p}) = \frac{1}{2\pi} \frac{\delta(p_1+p_2+p_3-1)}{\sqrt{p_1 p_2 p_3}} \delta(p_1 - p_3) \tag{18}$$

and what follows. On marginalising away p_1 and p_3 (each equal to $\frac{1}{2}(1-p_2)$), we reach the posterior

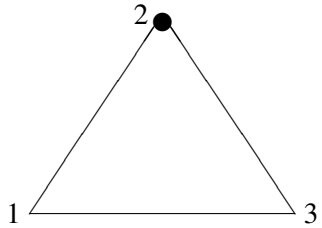
$$\Pr(p_2 \mid \text{data}) \propto \frac{1}{(1-p_2)\sqrt{p_2}} \qquad \text{⚠} \tag{19}$$

which has a non-integrable singularity at $p_2 = 1$.

With probability 1, p_2 is inferred to be arbitrarily close to 1. On observing p_1 to be equal to p_3, we are thus invited to infer that both are arbitrarily close to zero. That, surely, over-interprets the observation. The Bayesian analysis is correct, so the informed user will reject the geometric prior (16).



What's expected



What's delivered

Second counter-example: Six faces

A 6-die, not known to be uniform, has proportions $p_1, p_2, p_3, p_4, p_5, p_6$ associated with its faces. The geometric prior is

$$\Pr(\mathbf{p}) = \frac{2}{\pi^3} \frac{\delta(p_1+p_2+p_3+p_4+p_5+p_6-1)}{\sqrt{p_1 p_2 p_3 p_4 p_5 p_6}} \tag{20}$$

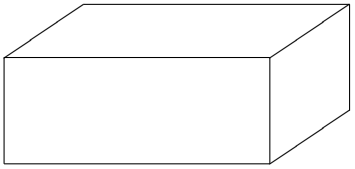
Accurate observation reveals that the die is a rectangular parallelepiped, with faces 1 and 6, 2 and 5, and 3 and 4, being equivalent. The likelihood is

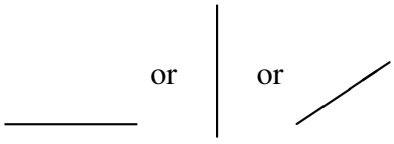
$$\Pr(\text{data} \mid \mathbf{p}) = \delta(p_1 - p_6)\delta(p_2 - p_5)\delta(p_3 - p_4) \tag{21}$$

On marginalising away p_4, p_5, p_6 away from the joint distribition, we reach

$$\Pr(p_1, p_2, p_3 \mid \text{data}) \propto \frac{1}{p_1 p_2 p_3} \qquad \text{⚠} \tag{22}$$

There are now three non-integrable singularities. With probability 1, only one component survives, either $p_1 = p_6$ or $p_2 = p_5$ or $p_3 = p_4$. The others are almost certainly almost zero. In lay terms, "all bricks are needles". Informed users would doubt that.


What's expected



What's delivered

Third counter-example: N items

The geometric prior for N items is

$$\Pr(\mathbf{p}) = \frac{\Gamma(\frac{N}{2})}{\pi^{N/2}} \frac{\delta(p_1+p_2+\ldots+p_N-1)}{\sqrt{p_1 p_2 \cdots p_N}} \tag{23}$$

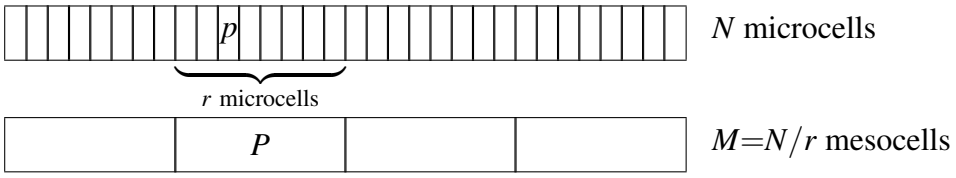
Measuring any accurate linear relationship $|ap_i - bp_j| < \varepsilon$ ensures that p_i and p_j will be inferred to be almost certainly arbitrarily close to zero as the uncertainty ε becomes small.

$$\underbrace{ap_i = bp_j}_{\text{data}} \qquad \Longrightarrow \qquad \underbrace{p_i = p_j = 0}_{\text{implication}} \qquad \text{⚠} \tag{24}$$

This over-implication needs no further comment.

Third counter-example: Continuum analysis

Suppose a continuum distribution is digitised into N microcells, with N large in order to approximate the continuum well. However, data are never infinitely sharp, so that it always suffices to combine the microcells, r at a time, into larger mesocells.



Marginalising (23) over r microcell p's summing to a mesocell's quantity P shows the mesocell prior to be

$$\Pr(P) \propto \iint \ldots \int \frac{\delta(p_1+p_2+\ldots+p_r-P)}{\sqrt{p_1 p_2 \cdots p_r}} dp_1 dp_2 \ldots dp_r \propto P^{-1+r/2} \tag{25}$$

so that the overall prior becomes

$$\Pr(\mathbf{P}) \propto (P_1 P_2 \ldots P_M)^{-1+r/2} \delta(P_1+P_2+\ldots+P_M-1) \qquad \text{⚠} \tag{26}$$

The exponents, which were $-1+\frac{1}{2}$ at the microscale, have become $-1+\frac{r}{2}$ at the mesoscale.

Now, what *ought* to happen as the continuum limit $r \to \infty$ is approached, is that the microscale exponent approaches -1 while mesoscale and macroscale exponents remain fixed. With power laws as here, this is the Dirichlet process [11], a "process" being a family of probability distributions defined consistently at all scales.

What is *actually* happening here is different. The microscale exponent is staying fixed at $-\frac{1}{2}$ while mesoscale and macroscale exponents increase indefinitely. This means that the prior for $\mathbf{p}$, at observable scales, becomes indefinitely sharply peaked about exact

uniformity ($P_1 = P_2 = \ldots = P_M = 1/M$). This contradicts the aim of allowing $\mathbf{p}$ to be usefully uncertain.

It *is* possible for $\mathbf{P}$ to be moved away from uniformity, but only by data that completely prohibit that possibility. In that event, $\mathbf{P}$ remains sharply defined, though relocated to the permitted maximum of $P_1 P_2 \ldots P_M$, equivalently of $\sum \log P_j$. But that's the Rényi-Tsallis prescription with $\alpha \to 0$, already seen to be unacceptable.

Conclusion regarding geometric densities

If the geometric $p^{-1/2}$ density is assigned at all, it has to be on a fixed grid, in which the cells can't be combined or subdivided. That grid can't be indefinitely fine, so that continuum problems are excluded. Even on a locked grid, unacceptable results follow accurate observation of any linear relationship.

Geometric manifolds

It seems unlikely that the difficulties remarked above would disappear when attention is restricted to a manifold within the probability simplex. If the density $\sqrt{\det g}$ fails in general, it's unlikely to succeed in arbitrary sub-spaces. Nevertheless, we investigate the possibility.

Parameters $\mathbf{u} = (u_1, u_2, \ldots)$, fewer in number than the dimension of the probability distribution, parameterise a manifold $\mathbf{p}(\mathbf{u})$ in a way that for convenience automatically imposes normalisation. The length element from (9), as confined to the manifold, becomes

$$(d\ell)^2 = \sum_i \frac{1}{p_i} \Big(\sum_j \frac{\partial p_i}{\partial u_j} du_j \Big) \Big(\sum_k \frac{\partial p_i}{\partial u_k} du_k \Big) = \sum_{jk} G_{jk} du_j du_k \tag{27}$$

where

$$G_{jk} = \sum_i \frac{1}{p_i} \frac{\partial p_i}{\partial u_j} \frac{\partial p_i}{\partial u_k} \tag{28}$$

is the metric tensor in the manifold. Consequently, the geometric density is

$$\rho(\mathbf{u}) \propto \sqrt{\det G} \tag{29}$$

Can this be used to assign prior probability over the manifold?

The simple answer is "generally no": it's dominated by the wrong properties. If the manifold allows large gradients $\partial p / \partial u$ to appear anywhere, then the density will be large and prior probability will coalesce there. Yet it's the *magnitude* of p that matters in probabilistic analysis, *not the gradient*. Local gradients tend to be unobservable because data have finite resolution, so they should surely not dominate the analysis. The supposition is fundamentally misdirected.

Counter-example: Growth and decay

A user seeks two locations around the unit circle. These are the minimum and maximum of a periodic distribution. The allowed distributions are functions $p(x)$ over the periodic unit interval $x \in [0,1)$, parameterised by the location u_1 of the minimum value, and the subsequent location u_2 of the maximum. From u_1 to u_2, the function p grows as

$$p(x) = f\Big(\frac{x-u_1}{u_2-u_1}\Big) \tag{30}$$

with a given monotonically increasing profile f. The same profile is used in reverse as

$$p(x) = f\Big(\frac{1+u_1-x}{1+u_1-u_2}\Big) \tag{31}$$

to give decay from u_2 to the next minimum at $1+u_1$. The profile f

(a) is normalised $\int_0^1 f(\theta)d\theta = 1$ to ensure normalisation of p;
(b) is strictly positive $f > 0$ to avoid division by zero;
(c) is differentiable with $f' > 0$ between its end points 0 and 1;
(d) has zero slope $f' = 0$ at those end points to avoid concern about matching.

Direct evaluation gives

$$\begin{pmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{pmatrix} = \frac{1}{(u_2-u_1)(1+u_1-u_2)} \begin{pmatrix} A & B \\ B & C \end{pmatrix} \tag{32}$$

where A,B,C are positive constants. For example, $f(t) = (8+6t^2-4t^3)/9$ gives $A = 0.01762, B = 0.01273, C = 0.01636$. This gives density

$$\rho(u_1,u_2) \propto \sqrt{\det G} = \frac{\sqrt{AC-B^2}}{(u_2-u_1)(1+u_1-u_2)} \qquad \text{⚠} \tag{33}$$

which is not normalisable, so cannot be used as a prior probability. The proposal fails.

If the attempt is nevertheless made, then with probability one either growth is instantaneous ($u_1 = u_2$) or decay is instantaneous ($u_2 = 1+u_1$). That's not what the user will have wanted. An ecologist interested in annual cycles would view askance the suggestion that either spring or autumn were instantaneous transitions between highest summer and deepest winter.

Geometry in thermodynamics

Physicists model systems by listing the allowed states, endowed with an appropriate counting measure (usually uniform, 1 per state). Each state i has associated observable "coordinates" $X_i^{(1)}, X_i^{(2)}, \ldots$ such as energy, volume,.... The system is to occupy its state subject to constraints on those values. Those values could in principle be known exactly, but it's more illuminating — and realistic — to constrain only average values so that the

occupancy is somewhat uncertain, being defined by a probability distribution $\mathbf{p}$ restricted by

$$\sum_i p_i X_i^{(k)} = \langle X^{(k)} \rangle = \text{fixed, for } k = 1, 2, \ldots. \tag{34}$$

Rational assignment of $\mathbf{p}$ is then uniquely defined by minimising $H(\mathbf{p}; \texttt{uniform})$ subject to the constraints, which produces the Gibbs distribution

$$p_i = Z^{-1} e^{-\sum_k \lambda_k X_i^{(k)}} \tag{35}$$

in which the "partition function"

$$Z(\lambda) = \sum_i e^{-\sum_k \lambda_k X_i^{(k)}} \tag{36}$$

ensures the normalisation $\sum_i p_i = 1$ that must always hold.

If we call the X's coordinates, we can equally call the Lagrange multipliers λ "forces". They control physical observables, so are themselves observable and carry physical interpretations such as coolness (inverse temperature) to control energy, pressure to control volume, and so on. The partition function encapsulates a neat summary of all this, as its derivatives

$$\frac{\partial \log Z}{\partial \lambda_k} = -\sum_i p_i X_i^{(k)} = -\langle X^{(k)} \rangle \tag{37}$$

are identifiable with the required constraint values. Going further, the second derivatives

$$\frac{\partial^2 \log Z}{\partial \lambda_k \partial \lambda_l} = \left\langle \left(X^{(k)} - \langle X^{(k)} \rangle \right) \left(X^{(l)} - \langle X^{(l)} \rangle \right) \right\rangle \tag{38}$$

identify the uncertainty covariance of the X's around their mean values. This uncertainty will manifest as observable fluctuations if their timescale isn't too long.

The Gibbs distribution (35) can be viewed either as a function of the constraints $\langle X \rangle$ or as a function of the λ's. Taking the latter view, the distributions form a manifold parameterised by λ, on which the metric (28) would evaluate to

$$G_{kl} = \sum_i p_i \left(X_i^{(k)} - \langle X_i^{(k)} \rangle \right) \left(X_i^{(l)} - \langle X_i^{(l)} \rangle \right) = \left\langle \left(X^{(k)} - \langle X^{(k)} \rangle \right) \left(X^{(l)} - \langle X^{(l)} \rangle \right) \right\rangle \tag{39}$$

This happens to be the same as (38) so that the geometric length element would be simply

$$(d\ell)^2 = \sum_{kl} \frac{\partial^2 \log Z}{\partial \lambda_k \partial \lambda_l} \, d\lambda_k d\lambda_l \tag{40}$$

The identification [12] is neat, but does it correspond to useful physics?

Example: Independent particles

In this simple example, the sole constraining coordinate X is energy E, which has just two levels, $E = 0$ and $E = 1$. Each level can be occupied independently by any of n equivalent classical particles. Accordingly there are in all 2^n states, which can be grouped into energy levels $r = 0, 1, 2, \ldots, n$, with ${}^nC_r$ states having energy r.

$$\text{Level } r\text{:} \left\{ \begin{array}{lll} E=1 & \bullet\bullet\bullet & r \text{ particles} \\ E=0 & \bullet\bullet\bullet\bullet\bullet & n-r \text{ particles} \end{array} \right\} {}^nC_r \text{ states}$$

As a function of coolness λ, the Gibbs distribution is

$$p_r = Z^{-1} \frac{n!}{r!(n-r)!} e^{-\lambda r} \tag{41}$$

with partition function

$$Z = \sum_{r=0}^{n} \frac{n!}{r!(n-r)!} e^{-\lambda r} = (1 + e^{-\lambda})^n \tag{42}$$

Its first derivative gives

$$\texttt{mean } \langle E \rangle = -\frac{\partial \log Z}{\partial \lambda} = \frac{n}{e^{\lambda} + 1} \tag{43}$$

so that (plausibly) energy ranges from the ground state $\langle E \rangle = 0$ at infinite coolness (zero temperature) up to $\langle E \rangle = n/2$ with all states equally occupied at zero coolness (infinite temperature). The physics is behaving properly.

What about geometry?

The geometric length element from (40) is

$$d\ell = \frac{n^{1/2}}{2\cosh(\lambda/2)} \, d\lambda \tag{44}$$

which integrates to

$$\ell(\lambda) = n^{1/2} \arctan \sinh(\lambda/2) \tag{45}$$

Unit length corresponds to unit fluctuation, and the full path between $\lambda = 0$ and $\lambda = \infty$ has length $O(\sqrt{n})$:

$$\ell(\infty) - \ell(0) = \tfrac{\pi}{2} n^{1/2} \tag{46}$$

With only one degree of freedom λ, the geometric prior density obeys $\rho \propto d\ell/d\lambda$, so is

$$\rho(\lambda) = \pi^{-1} \operatorname{sech}(\lambda/2) \tag{47}$$

Low temperature (high λ) is exponentially improbable, which might disconcert the low-temperature physicist with a rather different prior expectation.

Is it the job of theory to dictate the domain of experimentation?

Counter-example: First-order phase change

There are again only two energy states, but internal attractions between the components dictate that the system resides either in the unique ground state $E = 0$ or in any of the e^n top states of energy $E = n$. This idealises a phase change between a cold condensed state ("water") and a hot gaseous phase ("steam") in which the n components all evaporate into a larger volume of high-energy states. As before, the size of the system is n.

$$E = n \quad \xrightarrow{\text{steam}} \quad e^n \text{ states}$$
$$E = 0 \quad \xrightarrow{\text{water}} \quad 1 \text{ state}$$

The Gibbs distribution covers the water state with no energy and e^n steam states with energy n, so is

$$p_{\text{water}} = Z^{-1}, \qquad p_{\text{steam}} = Z^{-1} e^n e^{-\lambda n}, \tag{48}$$

with partition function

$$Z = 1 + e^n e^{-\lambda n} \tag{49}$$

Its first derivative gives

$$\texttt{mean}\ \langle E \rangle = -\frac{\partial \log Z}{\partial \lambda} = \frac{n}{e^{(\lambda-1)n} + 1} \tag{50}$$

which again ranges from the ground state ($\lambda = \infty$) to equal-occupancy ($\lambda = 0$). The "boiling-point" transition at $\lambda = 1$ is sharp, with only a small interval $\delta\lambda \sim n^{-1}$ between almost all water and almost all steam. Recall that in macroscopic thermodynamics, n is large, of the order of Avogadro's number 10^{24}, so the boiling-point is *very* sharply defined. This physics is correct and understood.

What about geometry?

The geometric length element from (40) is

$$d\ell = \frac{n/2}{\cosh(\frac{1}{2}(\lambda-1)n)}\, d\lambda \tag{51}$$

which integrates to

$$\ell(\lambda) = \arctan\sinh(\tfrac{1}{2}(\lambda-1)n) \tag{52}$$

with total length

$$\ell(\infty) - \ell(0) \approx \pi \qquad \text{⚠} \tag{53}$$

This means that the transition, which involves macroscopic changes in energy and entropy, is only assigned $O(1)$ length, so that water and steam are only assigned comparatively minuscule geometric separation. In placing highly distinct states together, geometry fails to reflect practical physics.

Moreover, the geometric prior density (29) for λ is

$$\rho(\lambda) = \frac{n/2\pi}{\cosh(\frac{1}{2}(\lambda-1)n)} \tag{54}$$

If, as suggested, this were used as a prior probability, it would imply that λ was *a priori* known (far more accurately than is experimentally possible) to be almost exactly 1. Specifically,

$$\lambda = 1 \pm \frac{\pi}{n} \qquad \text{⚠} \tag{55}$$

(mean $\pm$ standard deviation). In other words, a macroscopic system is known to be at its transition temperature, just because a transition exists. That's obviously counter-factual.

OVERALL CONCLUSIONS

Information geometry promotes the information-based (Fisher-Rao) local metric to global status, thereby inducing macroscopic lengths and distances. That's mathematics but it's not science.

It is central and critical for science that independent systems are allowed to behave independently. The only connective that allows independence is H, which is "from-to" asymmetric so cannot be a distance. Geometry can be imposed mathematically, but conflicts with scientific expectation quickly appear in the very simplest of tests.

For use in science, theories should always, *always*, be appropriately tested, not just developed as mathematical formalism. Testing lies at the indispensible heart of scientific methodology, and underlies the reliable performance that is required there.

REFERENCES

1. Cox, R.T. 1946. Probability, frequency, and reasonable expectation. *Am. J. Phys.* **14**, 1–13.
2. Knuth, K.H. and Skilling, J. 2012. Foundations of Inference *Axioms* **1**, 38–73.
3. Shannon, C.E. 1951. A mathematical theory of communication *Bell Syst. Tech. J.* **27**, 379–423, 623–656.
4. Kullback, S. and Leibler, R.A. 1951. On information and sufficiency *Ann. Math. Stat.* **22**, 79–86.
5. Rényi, A. 1960. On measures of entropy and information *Proc. 4th Berkeley Symp. on Math. Statistics and Probability* **1**, 547–561.
6. Tsallis, C. 1988. Possible generalization of Boltzmann-Gibbs statistics *J. Statistical Physics* **52**, 479–487.
7. Amari, S. 1985. Differential-geometrical methods in statistics, *Lecture notes in statistics*, Springer-Verlag, Berlin.
8. Fisher, R. A. 1925. Theory of statistical estimation *Proc. Camb. Phil. Soc.* **122**, 700–725.
9. Rao, C. R. 1945. Information and the accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.* **37**, 81–89.
10. Skilling, J. 2006. Nested sampling for general Bayesian computation. *Bayesian Analysis* **1**, 833–860.
11. Bernardo, J.M. and Smith, A.F.M. 2000. *Bayesian theory*, John Wiley, London.
12. Crooks, G.E. 2007. Measuring thermodynamic length. *Phys. Rev. Lett*, **99**, 100602.