

Foundations of Inference

Kevin H. Knuth
Departments of Physics and Informatics
University at Albany (SUNY)
Albany NY 12222, USA

John Skilling
Maximum Entropy Data Consultants
Kenmare, Ireland

August 31, 2010

Abstract

We present a foundation for inference that unites and significantly extends the approaches of Kolmogorov and Cox. Our approach is based on quantifying finite lattices of logical statements in a way that satisfies general lattice symmetries. With other applications in mind, our derivations assume minimal symmetries, relying on neither complementarity nor continuity or differentiability. Each relevant symmetry corresponds to an axiom of quantification, and these axioms are used to derive a unique set of rules governing quantification of the lattice. These rules form the familiar probability calculus. We also derive a unique quantification of divergence and information. Taken together these results form a simple and clear foundation for the quantification of inference.

1 Introduction

The quality of an axiom rests on it being both *convincing* for the application(s) in mind, and *compelling* in that its denial would be intolerable.

We present elementary symmetries as convincing and compelling axioms for valuation (measure), bivaluation (probability), and divergence (information and entropy). Our aim is to provide a simple and widely comprehensible foundation for the standard quantification of inference. We make minimal assumptions — not just for aesthetic economy of hypotheses, but because simpler foundations have wider scope.

It is a remarkable fact that algebraic symmetries can imply a unique calculus of quantification. Section 2 lists the symmetries that are actually needed to derive the results, and writes each required symmetry as an axiom of quantification. In section 3, we derive the sum rule for valuation from the associative

symmetry of combination. This sum rule is the basis of measure theory. It's usually taken as axiomatic, but in fact it's derived from compelling symmetry, which explains its wide utility. There is also a direct-product rule for independent measures, again derived from associativity. Section 4 derives from the direct-product rule a unique quantitative divergence from source measure to destination.

In section 5 we derive the chain product rule for probability from the associativity of chained order (in inference, implication). Probability calculus is then complete. Finally, section 6 derives the Shannon entropy and information (*a.k.a.* Kullback-Leibler) as special cases of divergence of measures. All these formulas are uniquely defined by elementary symmetries alone, whose compelling relevance explains the widespread failure of alternative proposals.

Our approach is constructivist, and we avoid unnecessary formality that might unduly confine our readership. Sets and quantities are deliberately finite because we have never encountered infinite quantity, infinite precision, or infinite information in our scientific endeavors, and neither do we expect to do so. In practice, infinity is unobservable and never more than a convenient abstraction from the finite. Thus we hold that as a matter of principle it is methodologically proper to axiomatize finite systems before any optional passage towards infinity. Assuming infinity at the outset encourages extra doubt without actually yielding any practical advantage.

R.T. Cox [2] showed the way by deriving the unique laws of probability from logical systems having a mere three elementary “atomic” propositions. By extension, those same laws applied to Boolean systems with arbitrarily many atoms and ultimately, where appropriate, to well-defined infinite limits. However, Cox needed to assume continuity and differentiability to define the calculus to infinite precision. Instead, we use arbitrarily many atoms to define the calculus to arbitrarily fine precision. Avoiding infinity in this way yields results that are adequate for all practical application, while avoiding unobservable subtleties of the continuum.

Our approach unites and significantly extends the approaches of Kolmogorov [8] and Cox, to form a foundation for inference that yields not just probability calculus, but also the unique quantification of divergence and information.

2 Symmetries and axioms

The minimal structure we need is a set of identifiable “atoms” a, b, c, \dots . These atoms combine through a “join” operator \vee which joins two atoms (or compound elements) into a compound element $z = x \vee y$. Taken together, the atoms and compounds form a join-semilattice. One can extend this to a lattice by including the null element \perp consisting of no atoms.

In inference, the N atoms represent the most fundamental exclusive statements we can make about the states of the world (more precisely, of our model of it — we make no ontological claim). They are exclusive in the sense that no two can both be true. The lattice is the full Boolean lattice comprised of

the powerset of all 2^N combinations of atoms, with the lattice join operation \vee being the logical OR. Statements can equivalently be formulated in terms of sets of possible states, which results in an isomorphic Boolean lattice of sets ordered by set inclusion. Here the two perspectives of logic and sets, on which the Cox and Kolmogorov foundations are based, are united within the lattice-theoretic framework. This powerset comprises the “hypothesis space” of all possible statements that one can make about a particular model of the world. Meaningful statements exclude the null element, which in inference is the nothing-is-true absurdity. A Boolean lattice has a rich family of symmetries, any or all of which would be legitimate to assume for a calculus of inference.

Our aim is to quantify the lattice structure. In other applications, atoms may have inherent dependencies, in which case not all 2^N elements are allowed. So, with other applications potentially in mind, we select as **axioms** only the most general symmetries that are actually needed for quantification, and avoid those that are specific to Boolean powersets. Specifically, we do not assume commutativity (though it holds in inference) and we do not assume complementation (though NOT is a valid unary operation in inference).

2.1 Symmetries

We begin by specifying the symmetries on which our axioms are based.

The null element \perp has

$$\perp \vee x = x \tag{0}$$

($x \vee \perp = x$ need not be separately assumed) on the grounds that including nothing should do nothing. Join obeys strict order

$$x < x \vee y \tag{1}$$

for non-null y ($y < x \vee y$ need not be separately assumed) on the grounds that including something else (that’s disjoint) makes x bigger. Here the binary ordering relation represented generically by $<$ represents logical implication (\Rightarrow) between different statements, or equivalently, proper subset inclusion (\subset) in the powerset representation. Join also preserves order from the right and from the left

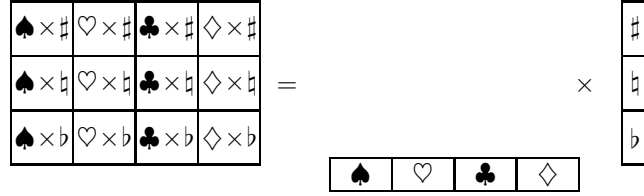
$$x \leq y \implies \begin{cases} x \vee z \leq y \vee z \\ z \vee x \leq z \vee y \end{cases} \tag{2}$$

for any z (a property that can be viewed as distributivity of \vee over \leq) on the grounds that ordering needs to be robust if it is to be useful. We do not assume either of the converses (cancellativity), on the grounds that in practical measurement the addition of a large number obscures small differences. Cancellativity and its cousin reducibility (where \leq is replaced by $=$) are not fully compelling for practical inference. Join is, however, assumed to be associative

$$(x \vee y) \vee z = x \vee (y \vee z) \tag{3}$$

on the grounds that this is a compelling symmetry of many systems, including inference.

Independent systems can be considered together. For example, one system might be playing-cards $x \in \{\spadesuit, \heartsuit, \clubsuit, \diamondsuit\}$, while another might be music keys $t \in \{b, \flat, \sharp\}$. The direct-product takes them both together, with atoms $x \times t$ like $\heartsuit \times \flat$.

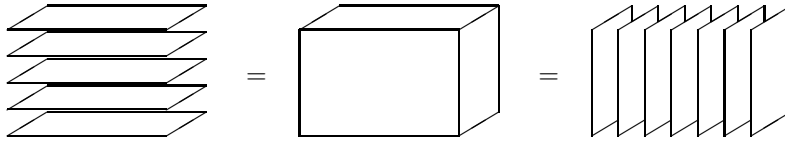


The direct-product operator \times is taken to be (right-)distributive over \vee

$$(x \times t) \vee (y \times t) = (x \vee y) \times t \quad (4)$$

on the grounds that relationships in one set, such as perhaps $\diamondsuit = \clubsuit \vee \heartsuit$, should remain intact whether or not an independent element from the other, such as perhaps \flat , is appended. Left distributivity is not needed. The direct-product of independent lattices is also taken to be associative

$$(u \times v) \times w = u \times (v \times w) \quad (5)$$



Again, we do not assume commutativity.

Finally, and in accordance with transitivity, concatenated relationships $\alpha = [x, y]$, $\beta = [y, z]$, $\gamma = [z, t]$ in a chain $x \leq y \leq z \leq t$ are associative

$$(\alpha, \beta), \gamma = \alpha, (\beta, \gamma) \quad (6)$$

These and these alone are the symmetries we need for the axioms of quantification. They are presented as a cartoon in the ‘‘Conclusions’’ section below.

2.2 Axioms

Our aim is to introduce a layer of quantification to the lattice. Our axioms arise from the requirement that any quantification must be consistent with the symmetries indicated above. Therefore, each symmetry gives rise to an axiom. We seek scalar valuations $m : \mathcal{L} \rightarrow \mathbb{R}$ to be assigned to elements x of a lattice \mathcal{L} , while conforming to the above structure (0),(1),...,(6) for disjoint elements.

The join operator $\vee : \mathcal{L} \times \mathcal{L} \rightarrow \mathcal{L}$ is quantified by the binary operator $\oplus : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ that determines $m(x \vee t) = m(x) \oplus m(t)$. In conformity with (0),

$$m(\perp) \oplus m(x) = m(x) \quad (\text{axiom 0})$$

To conform to (1), strict ordering, we require \oplus to obey

$$m(x) < m(x) \oplus m(y) \quad \text{(axiom 1)}$$

and its twin $m(y) \leq m(x) \oplus m(y)$ will eventually follow as consequence. To conform to (2), preservation of order, we require

$$m(x) \leq m(y) \implies \begin{cases} m(x) \oplus m(z) \leq m(y) \oplus m(z) \\ m(z) \oplus m(x) \leq m(z) \oplus m(y) \end{cases} \quad \text{(axiom 2)}$$

To conform to associativity (3), we require

$$(m(x) \oplus m(y)) \oplus m(z) = m(x) \oplus (m(y) \oplus m(z)) \quad \text{(axiom 3)}$$

These equations are to hold for arbitrary values $m(x)$, $m(y)$, $m(z)$ assigned to the disjoint x , y , z . Appendix A will show that these four axioms are necessary and sufficient to determine the calculus of measure.

The direct-product operator $\times : \mathcal{L} \times \mathcal{L}' \rightarrow \mathcal{L}''$ is quantified by the binary operator $\otimes : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ that determines $m(x \times t) = m(x) \otimes m(t)$. To conform to distributivity (4), we require

$$m(x \times t) \oplus m(y \times t) = m((x \vee y) \times t) \quad \text{(axiom 4)}$$

for disjoint x and y combined with any t from the second lattice. Presence of t may change the measures, but does not change their underlying additivity. Conforming to associativity (5) requires

$$(m(u) \otimes m(v)) \otimes m(w) = m(u) \otimes (m(v) \otimes m(w)) \quad \text{(axiom 5)}$$

These axioms will lead to a unique divergence between measures.

Finally, pair relationships like $\zeta = [x, t]$ need a bivaluation $p : \mathcal{L} \times \mathcal{L} \rightarrow \mathbb{R}$ that we could write as $p(\zeta)$ but is usually written as $p(x | t)$. Concatenation of pair relationships along a chain is quantified by a binary operator $\odot : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ obeying associativity (6)

$$(p(\alpha) \odot p(\beta)) \odot p(\gamma) = p(\alpha) \odot (p(\beta) \odot p(\gamma)) \quad \text{(axiom 6)}$$

when $\alpha = [x, y]$, $\beta = [y, z]$, $\gamma = [z, t]$ concatenate as the chain $x \leq y \leq z \leq t$. This final axiom will let us pass from measure to probability and Bayes' theorem, and from divergence to information and entropy.

| Operator | elements | quantification |
|----------------|----------|----------------|
| ordering | < | < |
| OR | \vee | \oplus |
| direct product | \times | \otimes |
| concatenation | , | \odot |

3 Measure

3.1 Disjoint arguments

According to the scalar *associativity theorem* (appendix A), an operator \oplus obeying axioms 0, 1, 2, 3 can without loss of generality be taken to be addition $+$, so that

$$m(x \vee y) = m(x) + m(y) \quad \text{(sum rule)}$$

In this form, valuation is known as a *measure*.

Commutativity of measure, $m(x \vee y) = m(y \vee x)$, though not explicitly assumed, follows as an unsurprising consequence. We also have $m(\perp) = 0$. Whilst we are free to adopt additivity as a convenient convention, we are also free to adopt any 1:1 regrading $m = \Theta(\mu)$ for which the rule would be

$$\mu(x \vee y) = \Theta^{-1}(\Theta(\mu(x)) + \Theta(\mu(y)))$$

This carries no extra generality because μ can be reverted to additivity by applying Θ , but we need such alternative grading later to avoid inconsistency between different assignments. There is no other freedom. If the linear form of sum rule is to be maintained, the only freedom is linear rescaling $\mu(x) = Cm(x)$.

Measure theory (see for example Halmos [6]) is usually introduced with additivity (countably additive or σ -additive) and non-negativity as “obvious” basic assumptions, with emphasis on the technical control of infinity in unbounded applications. Here we emphasize the foundation, and discover the *reason* why measure theory is constructed as it is. The symmetry of a lattice requires it. Any other formulation would break basic symmetries, and would not yield a widely useful theory.

3.2 Arbitrary arguments

For elements x and y that need not be disjoint, their join is defined as comprising all their constituent atoms counted once only, and the meet \wedge as comprised of those atoms they have in common. In inference, \vee is logical OR and \wedge is logical AND.

By putting $x = u \vee v$ and $y = v \vee w$ for disjoint u, v, w , we reach the general “inclusion/exclusion” sum rule for arbitrary x and y

$$m(x \vee y) + m(x \wedge y) = m(x) + m(y)$$

Commutativity of \wedge ,

$$m(x \wedge y) = m(y \wedge x)$$

follows from the already-known commutativity $m(x \vee y) = m(y \vee x)$ of \vee .

3.3 Independence

From axiom 5, the associativity theorem (appendix A again) requires an additivity relationship that in general reads

$$\Theta(m(x \times t)) = \Theta(m(x)) + \Theta(m(t))$$

for some invertible function Θ of the measures m . We can't proceed as before to re-grade in terms of $\Theta(m)$ to supersede m , because we are already using additivity

$$m(x \times t) + m(y \times t) = m((x \vee y) \times t)$$

(axiom 4, distributivity of \times) to define the grade of m . Instead, we require consistency with that sum-rule behavior for elements $x \times t$ and $y \times t$. Defining $\Psi = \Theta^{-1}$ gives, term by term,

$$\Psi(\xi + \tau) + \Psi(\eta + \tau) = \Psi(\zeta(\xi, \eta) + \tau)$$

where

$$\xi = \Theta(m(x)), \quad \eta = \Theta(m(y)), \quad \zeta = \Theta(m(x \vee y)), \quad \tau = \Theta(m(t)).$$

Among these variables, ξ, η, τ are independent, but (through the sum rule), ζ depends on ξ and η but not τ . This is the *product equation*.

The product theorem (appendix B) shows Θ to be logarithm, so that \otimes was multiplication and

$$m(x \times t) = m(x) m(t) / C$$

in which m (known to be positive) takes the sign of the arbitrary constant C (which must also be positive). The obvious convention $C = 1$ loses no generality, and gives

$$m(x \times t) = m(x) m(t) \quad \text{(direct-product rule)}$$

Measures are required to multiply, because of associativity of direct product, and the “ $\times t$ ” operation simply means “scale by $m(t)$ ”. This is consistent with linear (t -dependent) rescaling being the only allowed freedom for the measure over x .

4 Variation

Variational principles are common in science and we seek one for measures. We seek a variational potential $H(\mathbf{m})$ whose constrained extremum allows the atom valuations $\mathbf{m} = (m_1, m_2, \dots)$ to be assigned subject to appropriate constraints. (Bold-face vector notation is used in this section.)

One scenario is that each atom i has its own constraint function $\lambda_i(m_i)$, in which case the variational equation is

$$\delta(H(\mathbf{m}) - \lambda_1(m_1) - \lambda_2(m_2) - \dots) = 0$$

Such separability is intrinsic to the variational approach, and requires an additive form

$$H(\mathbf{m}) = \sum_{\text{atoms } i} H(m_i)$$

from which perturbation of the i 'th valuation determines m_i through

$$H'(m_i) = \text{constraint functions}$$

where prime ($'$) indicates derivative.

Another scenario is direct product $x \times y$, where m_x is determined by constraints on x , and m_y by constraints on y , leaving the product $m_x m_y$ as the target value. Hence the variational assignment derives from

$$H'(m_x m_y) = \lambda(m_x) + \mu(m_y)$$

for constraint functions λ for x and μ for y . The variational theorem (appendix C) gives the solution of this functional equation as

$$H(m) = A + Bm + C(m \log m - m)$$

for the individual valuation being considered, where A, B, C are constants. We are allowed nonlinear dependence because H need not obey order: $x < y$ need not imply $H(x) \leq H(y)$.

The scaling of a variational potential is arbitrary (and can be absorbed in the constraint functions), so we may set $C = 1$, ensuring that H has a minimum rather than a maximum. Alternatively, $C = -1$ would ensure a maximum. However, the settings of A and B depend on the application.

Combining all the atoms yields

$$H(\mathbf{m}) = \sum_{\text{atoms } i} (A_i + B_i m_i + C(m_i \log m_i - m_i))$$

4.1 Divergence

One use of H is as a quantifier of the divergence of destination values \mathbf{w} from source values \mathbf{u} that existed before the constraints were applied. For this, we set $C = 1$ to get a minimum, $B_i = -\log u_i$ to place the unconstrained minimizing \mathbf{w} at \mathbf{u} , and $A_i = u_i$ to make the minimum value zero. This form is

$$H(\mathbf{w} | \mathbf{u}) = \sum_{\text{atoms } i} (u_i - w_i + w_i \log(w_i/u_i)) \quad \text{(divergence)}$$

This formula is unique even though the divergence it specifies is intrinsically non-commutative (“ \mathbf{w} from \mathbf{u} ” and “ \mathbf{u} from \mathbf{w} ” differ) so is not a true distance.

In the limit of many small values, H admits a continuum limit

$$H(\mathbf{w} | \mathbf{u}) = \int (u(\theta) - w(\theta) + w(\theta) \log(w(\theta)/u(\theta))) d\theta$$

The constraints that force a measure away from the original source may admit several destinations, but minimizing H is the unique rule that defines a defensibly optimal choice.

5 Probability Calculus

In inference, we seek to impose on the hypothesis space a quantified *degree of implication* $p(x | t)$, to represent the plausibility of x conditional on current knowledge that excludes all hypotheses outside the stated context t . It should depend on both x (obviously) and t (otherwise it would be just the measure of x). The natural conjecture is that probability should be identified with a normalized measure, and we proceed to prove this.

In general, we simply wish to set up a bivaluation for predicate x within context t .

5.1 Chained arguments

Within given context t , we require $p(x | t)$ to have the symmetries 0, 1, 2, 3 that define a measure. Consequently, p obeys the sum rule

$$p(x | t) + p(y | t) = p(x \vee y | t)$$

for disjoint x and y with $x \vee y \leq t$. It is the dependence on t that remains to be determined.

Association of concatenation (axiom 6) for a chain $x \leq y \leq z \leq t$ is represented by

$$\underbrace{(p(x | y) \odot p(y | z))}_{\alpha} \odot \underbrace{p(z | t)}_{\gamma} = \underbrace{p(x | y)}_{\alpha} \odot \underbrace{(p(y | z) \odot p(z | t))}_{\beta}$$

By the associativity theorem, there is a scale on which \odot is simple addition. However, we can't regrade to that scale and discard the original because we have already fixed the behaviour of p to be additive with respect to its first argument. Instead, we infer additivity on some other grade $\Theta(p)$

$$\Theta(\underbrace{p(x | z)}_{\alpha \odot \beta}) = \Theta(\underbrace{p(x | y)}_{\alpha}) + \Theta(\underbrace{p(y | z)}_{\beta})$$

required to be consistent with the sum-rule behaviour of p . Defining $\Psi = \Theta^{-1}$ gives

$$\underbrace{p(x | z)}_{\alpha \odot \beta} = \Psi(\Theta(\underbrace{p(x | y)}_{\alpha}) + \Theta(\underbrace{p(y | z)}_{\beta}))$$

Substituting this in the sum rule, term by term, yields the same *product equation* as before,

$$\Psi(\xi + \tau) + \Psi(\eta + \tau) = \Psi(\zeta(\xi, \eta) + \tau)$$

where

$$\xi = \Theta(p(x | z)), \quad \eta = \Theta(p(y | z)), \quad \zeta = \Theta(p(x \vee y | z)), \quad \tau = \Theta(p(z | t)).$$

Through the sum rule, ζ depends as shown on ξ and η but not τ . The independent variables are ξ, η, τ .

The solution (appendix B again) shows Θ to be logarithm, so that \odot was multiplication and

$$p(x | z) = p(x | y) p(y | z) / C$$

in which p (positive) takes the sign of a universal constant C . Without loss of generality, we assign the scale of p by fixing $C = 1$, giving the standard product rule for conditioning.

$$p(x | z) = p(x | y) p(y | z) \quad \text{(chain-product rule)}$$

The chain-product rule, which as written above is valid for any chain, can be generalized to accommodate arbitrary elements. This is accomplished by noting that $x \wedge y = x$ in a chain, so that $p(x \wedge y | y) = p(x | y)$. The general form

$$p(a \wedge b | c) = p(a | b \wedge c) p(b | c)$$

follows by observing that $x = a \wedge b \wedge c$, $y = b \wedge c$ and $z = c$ form a chain and hence are subject to the chain rule.

The special case $p(t | t) = 1$ is obtained by setting $y = z = t$ in the chain-product rule. For any $x \leq t$, ordering requires $p(x | t) \leq p(t | t) = 1$, so that the range of values is $0 \leq p \leq 1$ and we recognize p as *probability*, hereafter Pr.

5.2 Probability

Probability calculus is now proved:

| | |
|---|-----------------|
| $0 = \Pr(\perp t) < \Pr(x t) \leq \Pr(t t) = 1$ | (range) |
| $\Pr(x \vee y t) + \Pr(x \wedge y t) = \Pr(x t) + \Pr(y t)$ | (sum rule) |
| $\Pr(x \wedge y t) = \Pr(x y \wedge t) \Pr(y t)$ | (chain-product) |

From the commutativity $\Pr(x \wedge y | t) = \Pr(y \wedge x | t)$ associated with \wedge , we obtain Bayes' Theorem

$$\Pr(x | \theta \wedge t) \Pr(\theta | t) = \Pr(\theta | x \wedge t) \Pr(x | t)$$

which can be simplified by making the common context implicit and considering the two bivaluations as valuations

$$\underbrace{\Pr(x | \theta)}_{\text{Likelihood}} \underbrace{\Pr(\theta)}_{\text{Prior}} = \underbrace{\Pr(\theta | x)}_{\text{Posterior}} \underbrace{\Pr(x)}_{\text{Evidence}} \quad \parallel t$$

relating data x and parameter θ (context t understood).

5.3 Probability as a ratio

Probability calculus can be subsumed in the single expression

$$\Pr(x | t) = \frac{m(x \wedge t)}{m(t)} \quad \forall x \forall t$$

for probability as a ratio of measures. Thus the calculus of probability is nothing more than the elementary calculus of proportions. As anticipated, within its context t , a probability distribution is simply the *shape* of the confined measure, automatically normalized to unit mass.

This is, essentially, the original discredited frequentist definition (see von Mises [11]) of probability, as the ratio of number of successes to number of trials. However, it is here retrieved at an abstract level, which bypasses the catastrophic difficulties of literal frequentism when faced with isolated non-reproducible situations. Just as ordinary addition is forced for valuations and measures, so ordinary proportions are forced for probability calculus.

6 Information and Entropy

Here, we take special cases of the variational potential H , appropriate for probability distributions instead of arbitrary measures.

6.1 Information

Within a given context, probability is a measure, normalized to unit mass. The divergence H of destination probability \mathbf{p} from source probability \mathbf{q} then simplifies to

$$\boxed{H(\mathbf{p} \mid \mathbf{q}) = \sum_k p_k \log \frac{p_k}{q_k}} \quad \text{(information)}$$

This is the Kullback-Leibler formula [9], well-known in statistics. If the final destination is a fully determined state, with a single p equal to 1 while all the others are necessarily 0, then we have the extreme case

$$H(\mathbf{p} \mid \mathbf{q}) = -\log q_k \quad \text{when } p_k = 1.$$

This represents the information gained on acquiring knowledge k — equivalently the surprise at finding k instead of any available alternative.

In the limit of many small values, H admits a continuum limit

$$H(\mathbf{p} \mid \mathbf{q}) = \int p(x) \log \frac{p(x)}{q(x)} dx$$

sometimes (with a minus sign) known as the cross-entropy.

6.2 Entropy

The variational potential

$$H(\mathbf{p}) = \sum_k (A_k + B_k p_k + C(p_k \log p_k - p_k))$$

can also quantify uncertainty. For this, we require zero uncertainty when one probability value equals to 1 (definitely present) and all the others are necessarily

0 (definitely not present). This is accomplished by setting $A_k = 0$ and $B_k = C$. Setting $C = -1$ gives the conventional scale, and yields

$$\boxed{S(\mathbf{p}) = - \sum_k p_k \log p_k} \quad (\text{entropy})$$

We call this “entropy”, and give it a separate symbol S as well as a separate name, to distinguish it from the previous “information” special case of divergence.

Entropy happens to be the expectation value of the information gained by deciding on one particular cell instead of any of the others in a partition.

$$S(\mathbf{p}) = \langle -\log p_k \rangle_k$$

It is a function of the partitioning as well as the probability distribution, which is why it doesn’t have a continuum limit. Plausibly, entropy has the following three properties:

- S is a continuous function of its arguments.
- If there are n equal choices, so that $p_k = 1/n$, then S is monotonically increasing in n .
- If a choice is broken down into subsidiary choices, then S adds according to probabilistic expectation, meaning $S(p_1, p_2, p_3) = S(p_1, p_2+p_3) + (p_2+p_3)S(p_2, p_3)$.

These are the three properties from which Shannon [10] originally proved the entropy formula. Here, we see that those properties, like that formula, are inevitable consequences of seeking a variational quantity for probabilities.

Information and entropy are near synonyms, and are often used interchangeably. As seen here, though, entropy S is different from H . It is a property of just one partitioned probability distribution, it has a maximum not a minimum, and it does *not* have a continuum limit. Its least value, attained when a single probability is 1 and all the others are 0, is zero. However, its value generally diverges as the partitioning deepens.

7 Conclusions

7.1 Summary

We start with a set $\{a, b, c, \dots\}$ of “atomic” elements which in inference represent the most fundamental exclusive statements we can make about the states (of our model) of the world. This is expanded by combination \vee to a set $\{x \vee y \vee \dots\}$, which in inference is a Boolean lattice called the hypothesis space of statements. This structure has rich symmetry, but other applications may have less and we select only what we need. The minimal assumptions are so simple that they can be drawn as the following cartoon.

| | |
|------------------------------|---|
| 0: \perp does nothing | |
| 1: \vee obeys strict order | |
| 2: \vee preserves order | |
| 3: \vee is associative | $\underbrace{(\clubsuit \vee \heartsuit)} \vee \spadesuit = \clubsuit \vee \underbrace{(\heartsuit \vee \spadesuit)}$ |
| | Measure |
| 4: \times is distributive | |
| 5: \times is associative | $\underbrace{(\clubsuit \times \heartsuit)} \times \spadesuit = \clubsuit \times \underbrace{(\heartsuit \times \spadesuit)}$ |
| | Divergence |
| 6: order is associative | |
| | Measure \longrightarrow Probability and Bayes Divergence \longrightarrow Information and Entropy |

Axiom 0 just expresses an inherent property (including nothing does nothing) of what's meant by combining elements. Axioms 1 and 2 represent minimal properties that are required of the combination operator \vee as it relates to ordering. Axiom 3 says that valuation must conform to the associativity of \vee . These axioms are trivially required in inference. By the associativity theorem (appendix A — see the latter part for a proof of minimality) they require the valuation to be a measure $m(x)$, with \vee represented by addition $+$ (the *sum rule*). Any 1:1 regrading is allowed, but such change alters no content so that the standard linearity can be adopted by convention. This is the rationale behind measure theory.

The direct product operator \times that represents independence is distributive (axiom 4) and associative (axiom 5), and consequently independent measures multiply (the *direct-product rule*). There is then a unique form of variational potential for assigning measures under constraints, yielding a unique divergence

of one measure from another.

Probability $\Pr(x \mid t)$ is to be a bivaluation, automatically a measure over predicate x within any specified context t . Axiom 6 expresses associativity of ordering relations (in inference, implications) and leads to the *chain-product rule* which completes probability calculus. The variational potential defines the information (Kullback-Leibler) carried by a destination probability relative to its source, and also yields the Shannon entropy of a partitioned probability distribution.

7.2 Commentary

We have presented a foundation for inference that unites and significantly extends the approaches of Kolmogorov [8] and Cox [2], yielding not just probability calculus, but also the unique quantification of divergence and information. Our approach is based on quantifying finite lattices of logical statements in such a way that quantification satisfies specified lattice symmetries. This generalizes algebraic implication, or equivalently subset inclusion, to a calculus of degrees of implication. It is remarkable that the calculus is unique.

Our derivations have relied on a set of explicit axioms based on simple symmetries. In particular, we have made no use of complementarity (NOT), which in applications other than inference may well not be present. Neither have we assumed any additive or multiplicative behavior (as did Kolmogorov [8], de Finetti [3] and Dupre and Tipler [4]) — we find that sum and product rules follow from elementary symmetry.

At the cost of lengthening the proofs in the appendices, we have avoided assuming continuity or differentiability. Yet we remark that such infinitesimal properties ought not influence the calculus of inference. If they did, those infinitesimal properties would thereby have observable effects. But detecting whether or not a system is continuous at the infinitesimal scale would require infinite information, which is never available. So assuming continuity and differentiability, had that been demanded by the technicalities of mathematical proof (or by our own professional inadequacy), would in our view have been harmless. As it happens, each appendix touches on continuity, but the arguments are appropriately constructed to avoid the assumption, so any potential controversy (see for example Halpern [7]) over infinite sets and the rôle of the continuum disappears.

Other than 1:1 regrading, any deviation from the standard formulas must inevitably contradict the elementary symmetries that underlie them, so that popular but weaker justifications [3] in terms of decisions, loss functions, or monetary exchange can be discarded as unnecessary. Indeed, we hold generally that it is a tactical error to buttress a strong argument (like symmetry) with a weak argument (like betting, say). Doing that merely encourages a skeptic to sow confusion by negating the weak argument, thereby casting doubt on the main thesis through a false impression that the strong argument might be circumvented too.

Finally, the approach from basic symmetry is productive. Goyal and ourselves [5] have used just that approach to show why *quantum theory* is forced to use complex arithmetic. Long a deep mystery, the sum and product rules of complex arithmetic are now seen as inevitably necessary to describe the basic interactions of physics. Elementary symmetry thus brings measure, probability, information and fundamental physics together in a remarkably unified synergy.

Acknowledgements

The authors would like to thank Janos Aczél, Ariel Caticha, Julian Center, Philip Goyal, Steve Gull, Jeffrey Jewell, Vassilis Kaburlasos, Fabio Mendez, Carlos Rodríguez. KHK was supported in part by the College of Arts and Sciences and the College of Computing and Information of the University at Albany, NASA Applied Information Systems Research Program (NASA NNG06GI17G) and the NASA Applied Information Systems Technology Program (NASA - NNX07AD97A). JS was supported by Maximum Entropy Data Consultants Ltd.

Appendix A: Associativity theorem

We seek the operator \oplus that combines real valuations m of disjoint elements of a lattice, subject to axioms 0, 1, 2, 3. Valuations being initially arbitrary, we are free to set $m(\perp) = 0$ by convention, so that

$$0 \oplus u = u \quad \text{(axiom 0)}$$

The operator \oplus is assumed to obey strict order

$$u < u \oplus v \quad \text{(axiom 1)}$$

and assumed to preserve order

$$u \leq v \implies \begin{cases} u \oplus w \leq v \oplus w \\ w \oplus u \leq w \oplus v \end{cases} \quad \text{(axiom 2)}$$

for arbitrary values u, v, w of elements other than \perp . We seek to determine \oplus subject to associativity

$$(u \oplus v) \oplus w = u \oplus (v \oplus w) \quad \text{(axiom 3)}$$

Theorem: \oplus is addition $+$ over positive reals, or a regrade so that

$$u \oplus v = \Theta^{-1}(\Theta(u) + \Theta(v))$$

where $\Theta : \mathbb{R} \rightarrow \mathbb{R}$ is monotonic strictly increasing.

Proof:

Axiom 1, together with the convention axiom 0, immediately implies that all valuations (other than of \perp) are strictly positive. General results apply to special cases. Here, we take arbitrarily many “atoms” whose values are drawn from some subset $\{a, b, c, \dots\}$ of strictly positive reals.

$$\bullet \bullet \bullet \bullet \dots \bullet \bullet \bullet \dots \bullet \bullet \bullet \bullet \dots \dots$$

$$a \ a \ a \ a \ \dots \ b \ b \ b \ \dots \ c \ c \ c \ c \ \dots \dots$$

The corresponding lattice consists of elements constructed from various numbers of a 's then b 's then c 's and so on, with no assumption of commutativity.

Consider the elements that are constructed from atoms of value a only. By associativity (axiom 3), the valuation of an element comprising n such atoms

$$\underbrace{a \oplus a \oplus \dots \oplus a}_n = m(n \text{ of } a)$$

is independent of how the constituent atoms are bracketed

$$\dots (a \oplus a) \oplus a \dots = \dots a \oplus (a \oplus a) \dots$$

in the construction tree, so can be expressed as a function m of cardinality n and atom value a , with

$$m(r \text{ of } a) \oplus m(s \text{ of } a) = m(r+s \text{ of } a).$$

By axiom 1, $m(r \text{ of } a) < m(r+s \text{ of } a)$, and consequently

$$m(\perp) \equiv m(0 \text{ of } a) < m(1 \text{ of } a) < \dots < m(r \text{ of } a) < m(r+1 \text{ of } a) < \dots$$

with $m(1 \text{ of } a) = a$.

Provided we take $a > 0$, this sequence is monotonic strictly increasing in correspondence with the linear assignment

$$m(r \text{ of } a) = ra$$

which without loss of generality we may adopt by convention. It obeys all four axioms and the only freedom that does not destroy ordering is monotonic strictly increasing regrade Θ . This convention already defines the basic additive scale, and has

$$ra \oplus sa = ra + sa$$

Now take the next atomic value b , and consider the subset of elements that use only it. These values either coincide with or interleave the existing multiples of a . Moreover, the interleaving is linearly consistent. For example, we can't have started with $m(2 \text{ of } b) \leq m(3 \text{ of } a)$ (suggesting $b/a \leq 3/2 = 1.5$), whilst also having $m(3 \text{ of } b) \geq m(5 \text{ of } a)$ (suggesting $b/a \geq 5/3 = 1.66\dots$). This is because

$$\underbrace{m(9 \text{ of } a)}_{aaaaaaaaa} \geq \underbrace{m(6 \text{ of } a) \oplus m(2 \text{ of } b)}_{aaaaaabb} \geq \underbrace{m(3 \text{ of } a) \oplus m(4 \text{ of } b)}_{aaabbbb} \geq \underbrace{m(6 \text{ of } b)}_{bbbbbb}$$

At each step, the last 3 a 's are replaced by 2 b 's. We do not assume commutativity, but we do acknowledge at each step that \vee preserves order (axiom 2) to justify $m(\dots aaa \dots) \geq m(\dots bb \dots)$. Similarly, on replacing 5 a 's by 3 b 's,

$$\underbrace{m(10 \text{ of } a)}_{aaaaaaaa} \leq \underbrace{m(5 \text{ of } a) \oplus m(3 \text{ of } b)}_{aaaaabbb} \leq \underbrace{m(6 \text{ of } b)}_{bbbbbb}$$

Hence $m(9 \text{ of } a) \geq m(10 \text{ of } a)$, which would have contradicted ordering.

Generally, “ $m(r_1 \text{ of } a) \leq m(s_1 \text{ of } b)$ ” and “ $m(r_2 \text{ of } a) \geq m(s_2 \text{ of } b)$ ” are only consistent if $r_1/s_1 \leq r_2/s_2$, which can be seen from

$$m(r_1 s_2 \text{ of } a) \leq m(s_1 s_2 \text{ of } b) \leq m(s_1 r_2 \text{ of } a)$$

as demonstrated above for $r_1 = 3$, $s_1 = 2$, $r_2 = 5$, $s_2 = 3$, which shows that

$$r_1 s_2 a \leq m(s_1 s_2 \text{ of } b) \leq s_1 r_2 a$$

Without contradicting the existing a -scale, we can adopt the convention

$$m(s \text{ of } b) = sb$$

provided the ratio b/a is assigned somewhere in the range

$$\frac{b}{a} \in \left[\max_{m(r \text{ of } a) \leq m(s \text{ of } b)} \left(\frac{r}{s} \right), \min_{m(r \text{ of } a) \geq m(s \text{ of } b)} \left(\frac{r}{s} \right) \right]$$

This obeys all four axioms and the only freedom that does not destroy ordering is monotonic strictly increasing regrade Θ . With all choices of r and s available, the maximum r beneath (\leq) and the minimum r above (\geq) any specified s will be adjacent or coincident integers. When s is taken indefinitely large — which can occur because in general a and b could be indefinitely small — the setting of b/a becomes arbitrarily precise, to within at worst $\pm 1/s$.

The same approach, with an extra r_0 a 's on the left, yields

$$m((r_0 + r_1 s_2) \text{ of } a) \leq m(r_0 \text{ of } a \text{ with } s_1 s_2 \text{ of } b) \leq m((r_0 + s_1 r_2) \text{ of } a)$$

which shows that

$$r_0 a + r_1 s_2 a \leq r_0 a \oplus s_1 s_2 b \leq r_0 a + s_1 r_2 a$$

Hence, to arbitrary precision,

$$\begin{aligned} m(r \text{ of } a \text{ with } s \text{ of } b) &\equiv r a \oplus s b \\ &= r a + s b \end{aligned}$$

This assignment obeys all four axioms and the only freedom that does not destroy ordering is monotonic strictly increasing regrade Θ .

For the next atomic value c , we similarly get bounds for c/a . With valuations of multiple b and multiple c both being linear with respect to multiple a , they are

linear with respect to each other, so there is no contradiction between b and c . We then have consistent valuations $ra + sb + tc$. By induction, we construct a global valuation

$$\begin{aligned} m(r \text{ of } a \text{ with } s \text{ of } b \text{ with } t \text{ of } c \text{ with } \dots) &\equiv ra \oplus sb \oplus tc \oplus \dots \\ &= ra + sb + tc + \dots \end{aligned}$$

in which \oplus is $+$. Once again, this assignment obeys all four axioms and the only freedom that does not destroy ordering is monotonic strictly increasing regrade Θ .

The special case where $r, s, t \dots$ are all 0 or 1 describes the general situation

$$\begin{array}{cccccccccccccccccccc} \bullet & \bullet & \bullet & \dots & \bullet & \bullet & \bullet & \downarrow & \bullet & \dots & \dots & \bullet & \bullet & \dots & \dots & \bullet & \bullet & \downarrow & \bullet & \bullet & \bullet & \downarrow & \bullet & \bullet & \bullet & \dots & \dots \\ a & a & a & & f & f & f & & f & & & h & h & & & q & q & q & & t & t & t & t & & & & & \end{array}$$

in which values may or may not happen to be equal:

$$\begin{aligned} m(1 \text{ of } f \text{ with } 1 \text{ of } q \text{ with } 1 \text{ of } t \text{ with } \dots) &\equiv f \oplus q \oplus t \oplus \dots \\ &= f + q + t + \dots \end{aligned}$$

Valuations being positive, this assignment clearly satisfies all four axioms. Hence we are always *allowed* to use ordinary addition $\oplus = +$, after which the *only* freedom is monotonic strictly increasing regrade Θ . \square

The scalar associativity equation is thus solved (by $+$) without any continuity or differentiability assumption, by just relying on the ordinary consistency of arithmetic on rational numbers. Addition happens to be a continuous and differentiable operation, but non-continuous and non-differentiable regrade is permitted.

Axioms are minimal

Theorem: Axioms 0, 1, 2, 3 are individually required.

Proof:

We construct operators \circ (“not quite \oplus ”) which deny each axiom in turn, while not being a monotonic strictly increasing regrade of addition.

Without axiom 0 (null does nothing), the definition

$$a \circ b = a + \lceil b \rceil$$

(where $\lceil b \rceil$ is the integer at or immediately above b) satisfies axioms 1, 2 and 3, but cannot be equivalent to addition because it’s non-commutative, $a \circ b \neq b \circ a$. So axiom 0 is required.

Without axiom 1 (strict ordering), the definition

$$a \circ b = \max(a, b)$$

satisfies axioms 0, 2 and 3, but is not equivalent to addition because it's insensitive to the lesser component. So axiom 1 is required.

To assess axiom 2 (preservation of order), take two combination rules

$$a \oplus b = a + b$$

$$p \underline{\oplus} q = \sqrt{p^2 + q^2}$$

each of which obeys all four axioms, but which are additive on different grades. Real numbers can be expressed in binary notation as

$$x = \sum_{-\infty}^{\text{top}} 2^i x_i = x_{\text{top}} \dots x_2 x_1 x_0 \cdot x_{-1} x_{-2} \dots$$

(e.g. $6\frac{1}{4} = 110.01$) and merged by interleaving their bit patterns

$$z = \text{merge}(x, y) = \sum 2^{2i+1} x_i + \sum 2^{2i} y_i = \dots x_1 y_1 x_0 y_0 \cdot x_{-1} y_{-1} x_{-2} y_{-2} \dots$$

From this merged number, the original parts can be recovered by extracting the appropriate (odd or even) bits, written as

$$x = \text{odd}(z), \quad y = \text{even}(z).$$

Define

$$x \circ y = \text{merge}(\text{odd}(x) \oplus \text{odd}(y), \text{even}(x) \underline{\oplus} \text{even}(y))$$

This satisfies axiom 0 trivially, axiom 1 by inheritance from the odd and even parts separately, and axiom 3 by similar inheritance. However, it does not satisfy axiom 2. Taking for example

$$u = 25 = \text{merge}(2, 5), \quad v = 67 = \text{merge}(1, 9), \quad w = 88 = \text{merge}(2, 12),$$

we have

$$u \circ w = \text{merge}(2 + 2, \sqrt{5^2 + 12^2}) = \text{merge}(4, 13) = 113,$$

$$v \circ w = \text{merge}(1 + 2, \sqrt{9^2 + 12^2}) = \text{merge}(3, 15) = 67.$$

so that

$$25 < 67 \quad \text{but} \quad 25 \circ 88 = 113 > 67 \circ 88 = 95.$$

This is inconsistent with monotonically regraded addition because " $\oplus 88$ " should preserve order whereas " $\circ 88$ " does not. So axiom 2 is required.

Without axiom 3 (associativity), the definition

$$a \circ b = a + b + ab^2$$

satisfies axioms 0, 1 and 2, but cannot be equivalent to addition because it's non-commutative, $a \circ b \neq b \circ a$. So axiom 3 is required too. \square

We do not claim these four to be the only minimal set. For example, Aczél [1] uses continuity and reducibility in place of our axioms 0, 1 and 2. Rather, we aim to provide a minimal set that is convincing and compelling for inference and other applications.

Appendix B: Product Theorem

Theorem: The solution of the functional equation

$$\Psi(\tau + \xi) + \Psi(\tau + \eta) = \Psi(\tau + \zeta(\xi, \eta)) \quad (\text{product equation})$$

in which τ , ξ and η are independent real variables is $\Psi(x) = Ce^{Ax}$ where A and C are constants.

Proof:

First, we take the special case $\xi = \eta$, so that $\zeta - \xi$ and $\zeta - \eta$ take a common value a . This gives a 2-term recurrence

$$2\Psi(\tau + \zeta - a) = \Psi(\tau + \zeta)$$

in which τ and ζ remain independent, though a might be constant. In fact, a must be constant otherwise there would be no solution for Ψ . Consequently, Ψ behaves geometrically with

$$\Psi(\theta + na) = 2^n \Psi(\theta)$$

for any integer n , θ being arbitrary. Although this plausibly suggests that Ψ will be exponential, that is not yet proved because Ψ could still be arbitrary within any assignment range of width a .

To complete the proof, take a second special case where $\zeta - \xi$ and $(\zeta - \eta)/2$ take a common value b . This gives a 3-term recurrence

$$\Psi(\tau + \zeta - b) + \Psi(\tau + \zeta - 2b) = \Psi(\tau + \zeta)$$

in which τ and ζ remain independent, though b might be constant. In fact, b must be constant otherwise there would be no solution for Ψ . The solution is

$$\Psi(\theta + mb) = \left(\frac{2\Psi(\theta)}{5+\sqrt{5}} + \frac{\Psi(\theta+b)}{\sqrt{5}} \right) \left(\frac{1+\sqrt{5}}{2} \right)^m + \left(\frac{2\Psi(\theta)}{5-\sqrt{5}} - \frac{\Psi(\theta+b)}{\sqrt{5}} \right) \left(\frac{-2}{1+\sqrt{5}} \right)^m$$

for any integer m , θ being arbitrary.

This combines with the 2-term formula to make

$$\begin{aligned} \Psi(\theta + mb - na) &= \left(\frac{2\Psi(\theta)}{5+\sqrt{5}} + \frac{\Psi(\theta+b)}{\sqrt{5}} \right) e^{m \log \left(\frac{1+\sqrt{5}}{2} \right) - n \log 2} + \\ &\quad (-1)^m \left(\frac{2\Psi(\theta)}{5-\sqrt{5}} - \frac{\Psi(\theta+b)}{\sqrt{5}} \right) e^{-m \log \left(\frac{1+\sqrt{5}}{2} \right) - n \log 2} \end{aligned}$$

For any integer n , there is an even integer m for which $0 \leq mb - na < 2b$ so that all three arguments of Ψ lie in the range $[\theta, \theta + 2b]$. As n is allowed to increase indefinitely, so does this m in proportion $m/n \approx a/b$. Depending on the sign of n , at least one of the exponents $\pm m \log \frac{1+\sqrt{5}}{2} - n \log 2$ can become indefinitely large and positive. Unbounded values of Ψ being unacceptable, the coefficient of that exponent must vanish. So either

$$\Psi(\theta + mb - na) = \Psi(\theta) e^{m \log \left(\frac{1+\sqrt{5}}{2} \right) - n \log 2}$$

(first term only) or

$$\Psi(\theta + mb - na) = (-1)^m \Psi(\theta) e^{-m \log\left(\frac{1+\sqrt{5}}{2}\right) - n \log 2}$$

(second term only, and even m makes the sign $(-1)^m = 1$). In the first case, bounded Ψ requires

$$\frac{b}{a} = \frac{\log\left(\frac{1+\sqrt{5}}{2}\right)}{\log 2}$$

and in the second case, bounded Ψ requires

$$\frac{b}{a} = -\frac{\log\left(\frac{1+\sqrt{5}}{2}\right)}{\log 2}$$

Either way,

$$\Psi(\theta + mb - na) = \Psi(\theta) e^{A(mb-na)}$$

with A constant.

Although this strongly suggests that Ψ will be exponential, that is not yet fully proved because offsets $mb - na$ with even m are only a subset of the reals. There could be one scaling for arguments θ of the form $mb - na$, another for the form $\sqrt{2} + mb - na$, yet another for $\pi + mb - na$, and so on. Fortunately, b/a is irrational, so the offset $mb - na$ can approach any real value x arbitrarily closely. Express x as $x = mb - na + \epsilon$ with m and n chosen to make ϵ arbitrarily small. Then

$$\Psi(x) = e^{A(mb-na)} \Psi(\epsilon) = e^{A(x-\epsilon)} \Psi(\epsilon) \approx e^{Ax} \Psi(\epsilon)$$

because $e^{A\epsilon} \approx 1$. Separating variables, $\Psi(\epsilon) \approx \text{constant}$, and

$$\Psi(x) = C e^{Ax} \tag{solution}$$

to arbitrarily high precision ($\epsilon \rightarrow 0$) with constant C .

This obeys the original product equation without further restriction, hence is the general solution, with corollary $e^{A\xi} + e^{A\eta} = e^{A\zeta}$ defining $\zeta(\xi, \eta)$ and confirming that $a = A^{-1} \log 2$ and $b = A^{-1} \log\left(\frac{1+\sqrt{5}}{2}\right)$ were appropriate constants. \square

The sought inverse, in terms of the constants A and C , is

$$\Theta(u) = \frac{1}{A} \log \frac{u}{C} \tag{inverse}$$

in which u takes the sign of C .

Appendix C: Variational Theorem

Theorem: The solution of the functional equation

$$H'(m_x m_y) = \lambda(m_x) + \mu(m_y) \tag{variational equation}$$

with positive m_x and m_y is

$$H(m) = A + Bm + C(m \log m - m)$$

where A, B, C are constants.

Proof:

Write $\log m_x = u$, $\log m_y = v$, and rewrite the functions as $\lambda^*(u)$, $\mu^*(v)$ and $H'(m) = h(\log m)$.

$$h(u + v) = \lambda^*(u) + \mu^*(v)$$

Put $v = 0$ to get $\lambda^*(u) = h(u) - \text{constant}$ and $u = 0$ to get $\mu^*(v) = h(v) - \text{constant}$.

$$h(u + v) = h(u) + h(v) - B$$

This is Cauchy's functional equation [1]

$$f(u + v) = f(u) + f(v)$$

for $f(t) = h(t) - B$ from which $f(nt) = nf(t)$ and then $f(\frac{r}{n}t) = \frac{r}{n}f(t)$ follow by induction for integer r and n . Hence

$$f(t) = ct$$

where $c = f(t_0)/t_0$ evaluated at any convenient base t_0 . Awkwardly, the recurrence only relates to a rational grid — there could be one value of c for rational multiples of 1, another value for rational multiples of $\sqrt{2}$, yet another for rational multiples of π , and so on. Fortunately, the sought function H is an integral of f , on which such infinitesimal detail has no effect.

To show that, we blur functions $\phi(u, v)$ by convolving them with the following unit-mass ellipse, chosen to blur u, v and $u+v$ equally, according to

$$\Phi(u, v) = \iint dx dy \frac{\mathbf{1}(x^2 + xy + y^2 < \frac{3}{4}\epsilon^2)}{\sqrt{3}\pi\epsilon^2/2} \phi(u - x, v - y)$$

For small width ϵ , blurring has negligible macroscopic effect. The convolution transforms the Cauchy equation to the same form

$$F(u + v) = F(u) + F(v)$$

as before, with the new function

$$F(t) = \int_{-\epsilon}^{\epsilon} dx \frac{2\sqrt{\epsilon^2 - x^2}}{\pi\epsilon^2} f(t - x)$$

being a continuous version of the original f , narrowly blurred over finite support. With continuity in place, the Cauchy solution

$$F(t) = Ct$$

can only have one value for the constant C .

Finally, the definition $dH/dm = h(\log m) = B + f(\log m)$ yields

$$\begin{aligned}
H(m) &= Bm + \int^m f(\log m') dm' && \text{(integrate)} \\
&= Bm + \int^{\log m} f(t) e^t dt && \text{(change variable)} \\
&= Bm + \int_{-\epsilon}^{\epsilon} dx \frac{2\sqrt{\epsilon^2 - x^2}}{\pi\epsilon^2} \int^{\log m} f(t) e^t dt && \text{(insert blurring)} \\
&= Bm + \int_{-\epsilon}^{\epsilon} dx \frac{2\sqrt{\epsilon^2 - x^2}}{\pi\epsilon^2} \int^{x+\log m} f(t-x) e^{t-x} dt && \text{(offset dummy } t) \\
&\approx Bm + \int_{-\epsilon}^{\epsilon} dx \frac{2\sqrt{\epsilon^2 - x^2}}{\pi\epsilon^2} \int^{\log m} f(t-x) e^t dt && (|x| \leq \epsilon \text{ small}) ! \\
&= Bm + \int^{\log m} F(t) e^t dt && \text{(definition of } F) \\
&= Bm + C \int^{\log m} t e^t dt && \text{(substitute)}
\end{aligned}$$

Hence, to arbitrarily high precision ($\epsilon \rightarrow 0$), H integrates to

$$H(m) = A + Bm + C(m \log m - m).$$

This obeys the original variational equation with corollaries $\lambda(x) = B_1 + C \log(x)$ and $\mu(x) = B_2 + C \log(x)$ where $B_1 + B_2 = B$, but with no further restriction, hence is the general solution. \square

References

- [1] J. Aczél, *Lectures on functional equations and their applications*, Academic Press, New York, 1966.
- [2] R. T. Cox, *Probability, frequency, and reasonable expectation*, Am. J. Physics **14** (1946), 1–13.
- [3] B. de Finetti, *Theory of probability, vol. i and vol. ii*, John Wiley and Sons, New York, 1974,1975.
- [4] M. J. Dupré and F. J. Tipler, *New axioms for rigorous bayesian probability*, Bayesian Anal. **4** (2009), no. 3, 599–606.
- [5] P. Goyal, K. H. Knuth, and J. Skilling, *Origin of complex quantum amplitudes and Feynman's rules*, Phys. Rev. A **81** (2010), 022109, arXiv:0907.0909 [quant-ph].
- [6] P. R. Halmos, *Measure theory*, Springer, Berlin, Heidelberg, New York, 1974.

- [7] J. Y. Halpern, *A counterexample to theorems of Cox and Fine*, *JAIR* **10** (1999), 67–85.
- [8] A. N. Kolmogorov, *Foundations of the theory of probability*, 2nd english ed., Chelsea, New York, 1956.
- [9] S. Kullback and R. A. Leibler, *On information and sufficiency*, *Ann. Math. Statist.* **22** (1951), no. 1, 79–86.
- [10] C. F. Shannon, *A mathematical theory of communication*, *Bell Syst. Tech. J.* **27** (1948), 379–423, 623–656.
- [11] R. von Mises, *Probability, statistics, and truth*, Dover, Mineola, NY, 1981, Republished from the second revised English edition published by George Allen and Unwin Ltd., London, in 1957.