Axioms **2012**, *1*, 38-73; doi:10.3390/axioms1010038

OPEN ACCESS

axioms

ISSN 2075-1680

www.mdpi.com/journal/axioms

Article

Foundations of Inference

Kevin H. Knuth [1,*] **and John Skilling** [2]

[1] Departments of Physics and Informatics, University at Albany (SUNY), Albany, NY 12222, USA

[2] Maximum Entropy Data Consultants Ltd., Kenmare, County Kerry, Ireland;
E-Mail: john@skilling.co.uk

* Author to whom correspondence should be addressed; E-Mail: kknuth@albany.edu;
Tel.: +1-518-442-4653; Fax: +1-518-442-5260.

Received: 5 March 2012; in revised form: 1 June 2012 / Accepted: 7 June 2012 / Published: 15 June 2012

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

Keywords: measure; divergence; probability; information; entropy; lattice

Classification: PACS 02.50.Cw

Classification: MSC 06A05

---

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, initially for measure, subsequently for probability, and finally for 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 gives the background and outlines the procedure and major results. Section 3 lists the symmetries that are actually needed to derive the results, and the following Section 4 writes each required symmetry as an axiom of quantification. In Section 5, we derive the sum rule for valuation from the associative symmetry of ordered combination. This sum rule is the basis of measure theory. It is usually taken as axiomatic, but in fact it is derived from compelling symmetry, which explains its wide utility. There is also a direct-product rule for independent measures, again derived from associativity. Section 6 derives from the direct-product rule a unique quantitative divergence from source measure to destination.

In Section 7 we derive the chain product rule for probability from the associativity of chained order (in inference, implication). Probability calculus is then complete. Finally, Section 8 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.

Our approach is constructivist, and we avoid unnecessary formality that might unduly confine our readership. Sets and quantities are deliberately finite since it is methodologically proper to axiomatize finite systems before any optional passage towards infinity. R.T. Cox [1] 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 cover all practical applications, while avoiding unobservable subtleties.

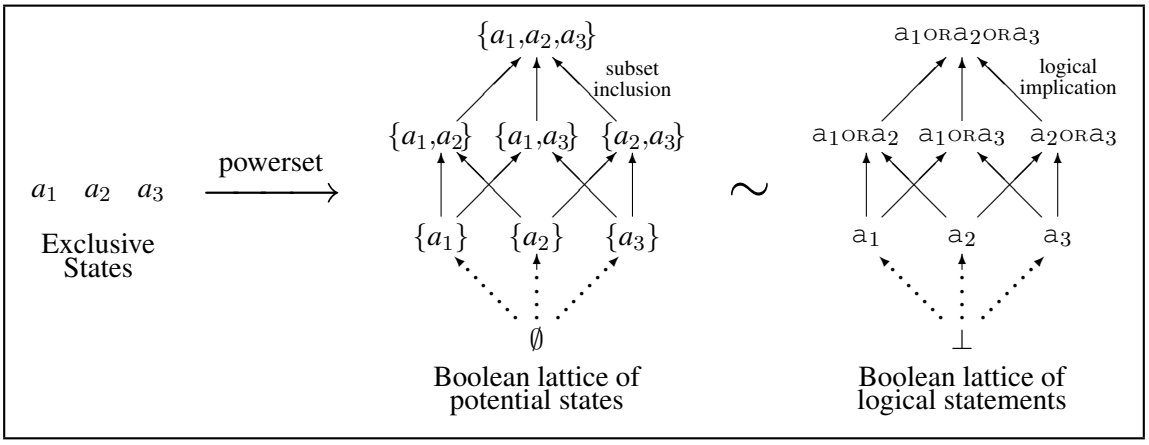
Our approach unites and significantly extends the set-based approach of Kolmogorov [2] and the logic-based approach of Cox [1], to form a foundation for inference that yields not just probability calculus, but also the unique quantification of divergence and information.

2. Setting the Scene

We model the world (or some interesting aspect of it) as being in a particular ***state*** out of a finite set of mutually exclusive states (as in Figure 1, left). Since we and our tools are finite, a finite set of states, albeit possibly very large in number, suffices for all practical modeling.

As applied to inference, each state of the world is associated, via isomorphism, with a statement about the world. This results in a set of mutually exclusive statements, which we call ***atoms***. Atoms are combined through logical `OR` to form compound statements comprising the ***elements*** of a ***Boolean lattice*** (Figure 1, right), which is isomorphic to a Boolean lattice of sets (Figure 1, center). Although carrying different interpretations, the mathematical structures are identical. Set inclusion "$\subset$" is equivalent to logical implication "$\Rightarrow$", which we abstract to lattice order "$<$". It is a matter of choice whether to include the null set $\emptyset$, equivalent to the logical absurdity $\perp$. The set-based view is ontological in character and associated with Kolmogorov, while the logic-based view is epistemological in character and associated with Cox.

Figure 1. The Boolean lattice of potential states (**center**) is constructed by taking the 2^N powerset of an antichain of N mutually exclusive atoms (in this case a_1, a_2, a_3, **left**). This lattice is isomorphic to the Boolean lattice of logical statements ordered by logical implication (**right**).



Quantification proceeds by assigning a real number $m(\mathtt{x}) = x$, called a ***valuation***, to elements $\mathtt{x}$. (`Typewriter` font denotes lattice elements $\mathtt{x}$, whereas their associated valuations (real numbers) x are shown in *italic*.) We require valuations to be faithful to the lattice, in the sense that

$$\underbrace{\mathtt{x} < \mathtt{y}}_{\text{lattice elements}} \quad \Longrightarrow \quad \underbrace{x < y}_{\text{real numbers}} \tag{1}$$

so that compound elements carry greater value than any of their components. Clearly, this by itself is only a weak restriction on the behavior of valuation.

Combination of two atoms (or disjoint compounds) into their compound is written as the operator $\sqcup$, for example $\mathtt{z} = \mathtt{x} \sqcup \mathtt{y}$. Our first step is to quantify the combination of disjoint elements through an operator $\oplus$ that combines values (Table 1 below lists such operators and their eventual identifications).

$$\underbrace{z = x \oplus y}_{\text{real numbers}} \quad \text{representing} \quad \underbrace{\mathtt{z} = \mathtt{x} \sqcup \mathtt{y}}_{\text{joined elements}} \tag{2}$$

We find that the symmetries underlying $\sqcup$ place constraints on $\oplus$ that effectively require it to be addition $+$. At this stage, we already have the foundation of ***measure theory***, and the generalization of combination (of disjoint elements) to the lattice join (of arbitrary elements) is straightforward. The wide applicability of these underlying symmetries explains the wide utility of measure theory, which might otherwise be mysterious.

Table 1. Operators and their symbols.

Operation	**Symbol**	**Quantification**	**(Eventual form)**
ordering	$<$	$<$	
combination	$\sqcup$	$\oplus$	(addition)
direct product	$\times$	$\otimes$	(multiplication)
chaining	,	$\odot$	(multiplication)

We can consider the atoms $\mathtt{a}_1, \mathtt{a}_2, \mathtt{a}_3, \ldots, \mathtt{a}_N$ and $\mathtt{b}_1, \mathtt{b}_2, \ldots, \mathtt{b}_M$ from separate problems as NM composite atoms $\mathtt{c}_{ij} = \mathtt{a}_i \times \mathtt{b}_j$ in an equivalent composite problem. The ***direct-product*** operator $\otimes$ quantifies the composition of values:

$$\underbrace{c = a \otimes b}_{\text{real numbers}} \qquad \text{representing} \qquad \underbrace{\mathtt{c} = \mathtt{a} \times \mathtt{b}}_{\text{composite element}} \tag{3}$$

We find that the symmetries of $\times$ place constraints on $\otimes$ that require it to be multiplication.

It is common in science to acquire numerical assignments by optimizing a variational potential. By requiring consistency with the numerical assignments of ordinary multiplication, we find that there is a unique variational potential $H(\mathbf{p} \mid \mathbf{q})$, of "$p \log p$" form, known as the (generalized Kullback–Leibler) Bregman ***divergence*** of measure $\mathbf{p}$ from measure $\mathbf{q}$.

Inference involves the relationship of one logical statement (predicate $\mathtt{x}$) to another (context $\mathtt{t}$), initially in a situation where $\mathtt{x} \Rightarrow \mathtt{t}$ so that the context includes subsidiary predicates. To quantify inference, we assign real numbers $p(\mathtt{x} \mid \mathtt{t})$, ultimately recognised as ***probability***, to predicate–context ***intervals*** $[\mathtt{x}, \mathtt{t}]$. Such intervals can be ***chained*** (concatenated) so that $[\mathtt{x}, \mathtt{z}] = [[\mathtt{x}, \mathtt{y}], [\mathtt{y}, \mathtt{z}]]$, with $\odot$ representing the chaining of values.

$$\underbrace{p(\mathtt{x} \mid \mathtt{z}) = p(\mathtt{x} \mid \mathtt{y}) \odot p(\mathtt{y} \mid \mathtt{z})}_{\text{real numbers}} \qquad \text{representing} \qquad \underbrace{[\mathtt{x}, \mathtt{z}] = [[\mathtt{x}, \mathtt{y}], [\mathtt{y}, \mathtt{z}]]}_{\text{chained intervals}} \tag{4}$$

We find that the symmetries of chaining require $\odot$ to be multiplication, yielding the ***product rule*** of probability calculus. When applied to probabilities, the divergence formula reduces to the ***information***, also known as the Kullback–Leibler formula, with ***entropy*** being a variant.

2.1. The Order-Theoretic Perspective

The approach we employ can be described in terms of order-preserving (monotonic) maps between order-theoretic structures. Here we present our approach, described above, from this different perspective.

Order-theoretically, a finite set of exclusive states can be represented as an ***antichain***, illustrated in Figure 1(left) as three states a_1, a_2, and a_3 situated side-by-side. Our state of knowledge about the world (more precisely, of our model of it—we make no ontological claim) is often incomplete so that we can at best say that the world is in one of a set of potential ***states***, which is a subset of the set of all possible states. In the case of total ignorance, the set of potential states includes all possible states. In contrast, perfect knowledge about our model is represented by singleton sets consisting of a single state. We refer to the singleton sets as ***atoms***, and note that they are exclusive in the sense that no two can be true.

The space of all possible sets of potential states is given by the partially-ordered set obtained from the powerset of the set of states ordered by set inclusion. For an antichain of mutually exclusive states, the powerset is a ***Boolean lattice*** (Figure 1, center), with the bottom element optional. By conceiving of a ***statement*** about our model of the world in terms of a set of potential states, we have an order-isomorphism from the Boolean lattice of potential states ordered by set inclusion to the Boolean lattice of statements ordered by logical implication (Figure 1, right). This isomorphism maps each set of

potential states to a statement, while mapping the algebraic operations of set union $\cup$ and set intersection $\cap$ to the logical OR and AND, respectively.

The perspective provided by order theory enables us to focus abstractly on the structure of a Boolean lattice with its generic algebraic operations ***join*** $\vee$ and ***meet*** $\wedge$. This immediately broadens the scope from Boolean to more general ***distributive lattices*** — the first fruit of our minimalist approach. For additional details on partially ordered sets and lattices in particular, we refer the interested reader to the classic text by Birkhoff [3] or the more recent text by Davey & Priestley [4].

Quantification proceeds by assigning valuations $m(\mathrm{x}) = x$ to elements x, to form a real-valued representation. For this to be faithful, we require an order-preserving (monotonic) map between the partial order of a distributive lattice and the total order of the chains that are to be found within. Thus $\mathrm{x} < \mathrm{y}$ is to imply that $x < y$, a relationship that we call ***fidelity***. The converse is not true: the total order imposed by quantification must be consistent with but can extend the partial order of the lattice structure.

We write the combination of two atoms into a compound element (and more generally any two disjoint compounds into a compound element) as $\sqcup$, for example $\mathrm{z} = \mathrm{x} \sqcup \mathrm{y}$. Derivation of the calculus of quantification starts with this disjoint combination operator, where we find that its symmetries place constraints on its representation $\oplus$ that allow us the convention of ordinary addition "$\oplus = +$". This basic result generalizes to the standard ***join*** lattice operator $\vee$ for elements that (possibly having atoms in common) need not be disjoint, for which the sum rule generalizes to its standard inclusion/exclusion form [5], which involves the meet $\wedge$ for any atoms in common.

There are two mathematical conventions concerning the handling the nothing-is-true null element $\perp$ at the bottom of the lattice known as the absurdity. Some mathematicians opt to include the bottom element on aesthetic grounds, whereas others opt to exclude it because of its paradoxical interpretation [4]. If it is included, its quantification is zero. Either way, fidelity ensures that other elements are quantified by positive values that are positive (or, by elementary generalization, zero). At this stage, we already have the foundation of ***measure theory***.

Logical deduction is traditionally based on a Boolean lattice and proceeds "upwards" along a chain (as in the arrows sketched in Figure 1). Given some statement x, one can deduce that x implies x OR y since x OR y includes x. Similarly, x AND y implies x since x includes x AND y. The ordering relationships among the elements of the lattice are encoded by the zeta function of the lattice [6]

$$\text{zeta function}: \qquad \zeta(\mathrm{x},\mathrm{y}) = \begin{cases} 1 & \text{if } \mathrm{x} \leq \mathrm{y} \\ 0 & \text{if } \mathrm{x} \not\leq \mathrm{y} \end{cases} \tag{5}$$

Deduction is definitive.

Inference, or ***logical induction***, is the inverse of deduction and proceeds "downwards" along a chain, losing logical certainty as knowledge fragments. Our aim is to quantify this loss of certainty, in the expectation of deriving probability calculus. This requires generalization of the binary zeta function $\zeta(\mathrm{x},\mathrm{y})$ to some real-valued function $p(x \mid y)$ which will turn out to be the standard probability of x GIVEN y. However, a firm foundation for inference must be devoid of a choice of arbitrary generalizations. By viewing quantification in terms of an order-preserving map between the partial order (Boolean lattice) and a total order (chain) subject to compelling symmetries alone, we obtain a firm foundation for inference, devoid of further assumptions of questionable merit.

By considering atoms (singleton sets, which are the join-irreducible elements of the Boolean lattice) as precise statements about exclusive states, and composite lattice elements (sets of several exclusive states) as less precise statements involving a degree of ignorance, the two perspectives of logic and sets, on which the Cox and Kolmogorov foundations are based, become united within the order-theoretic framework.

In summary, the powerset comprises the ***hypothesis space*** of all possible statements that one can make about a particular model of the world. Quantification of join using $+$ is the ***sum rule*** of probability calculus, and is required by adherence to the symmetries we list. It fixes the valuations assigned to composite elements in terms of valuations assigned to the atoms. Those latter valuations assigned to the atoms remain free, unconstrained by the calculus. That freedom allows the calculus to apply to inference in general, with the mathematically-arbitrary atom valuations being guided by insight into a particular application.

2.2. Commentary

Our results—the sum rule and divergence for measures, and the sum and product rules with information for probabilities—are standard and well known (their uniqueness perhaps less so). The matter we address here is which assumptions are necessary and which are not. A Boolean lattice, after all, is a special structure with special properties. Insofar as fewer properties are needed, we gain generality. Wider applicability may be of little value to those who focus solely on inference. Yet, by showing that the basic foundations of inference have wider scope, we can thereby offer extra—and simpler—guidance to the scientific community at large.

Even within inference, distributive problems may have relationships between their atoms such that not all combinations of states are allowed. Rather than extend a distributive lattice to Boolean by padding it with zeros, the tighter framework immediately empowers us to work with the original problem in its own right. Scientific problems (say, the propagation of particles, or the generation of proteins) are often heavily conditional, and it could well be inappropriate or confusing to go to a full Boolean lattice when a sparser structure is a more natural model.

We also confirm that commutativity is not a necessary assumption. Rather, commutativity of measure is imposed by the associativity and order required of a scalar representation. Conversely, systems that are not commutative (matrices under multiplication, for example) cannot be both associative and ordered.

3. Symmetries

Here, we list the relevant symmetries on which our axioms are based. All are properties of distributive lattices, and our descriptions are styled that way so that a reader wary of further generality does not need to move beyond this particular, and important, example. However, one may note that not all the properties of a distributive lattice (such as commutativity of the join) are listed, which implies that these results are applicable to a broader class of algebraic structures that includes distributive lattices.

Valuation assignments rank statements via an order-preserving map which we call ***fidelity*** .

$$\text{Symmetry 0}: \qquad \underbrace{\mathtt{x} < \mathtt{y}}_{\text{lattice elements}} \quad \Longrightarrow \quad \underbrace{x < y}_{\text{real numbers}} \tag{6}$$

It is a matter of convention that we choose to order the valuations in the same sense as the lattice order ("more is bigger"). Reverse order would be admissible and logically equivalent, though less convenient.

In the specific case of Boolean lattices of logical statements, the binary ordering relation, represented generically by $<$, is equivalent to logical implication ($\Rightarrow$) between *different* statements, or equivalently, proper subset inclusion ($\subset$) in the powerset representation. Combination preserves order from the right and from the left

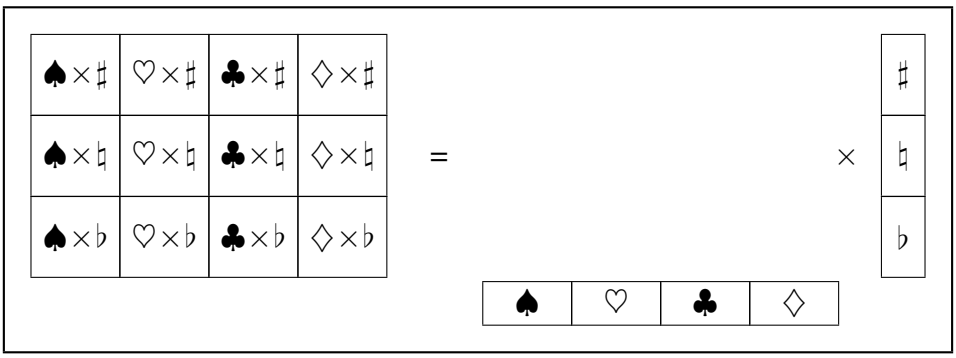
$$\text{Symmetry 1}: \qquad \mathtt{x} < \mathtt{y} \quad \Longrightarrow \quad \begin{cases} \mathtt{x} \sqcup \mathtt{z} < \mathtt{y} \sqcup \mathtt{z} \\ \mathtt{z} \sqcup \mathtt{x} < \mathtt{z} \sqcup \mathtt{y} \end{cases} \tag{7}$$

for any $\mathtt{z}$ (a property that can be viewed as distributivity of $\sqcup$ over $<$) on the grounds that ordering needs to be robust if it is to be useful. Combination is also taken to be associative

$$\text{Symmetry 2}: \qquad (\mathtt{x} \sqcup \mathtt{y}) \sqcup \mathtt{z} = \mathtt{x} \sqcup (\mathtt{y} \sqcup \mathtt{z}) \tag{8}$$

Independent systems can be considered together (Figure 2).

Figure 2. One system might, for example, be playing-card suits $\mathtt{x} \in \{\spadesuit, \heartsuit, \clubsuit, \diamondsuit\}$, while another independent system might be music keys $\mathtt{t} \in \{\flat, \natural, \sharp\}$. The direct-product combines the spaces of$\mathtt{x}$ and $\mathtt{t}$ to form the joint space of$\mathtt{x} \times \mathtt{t}$ with atoms like $\heartsuit \times \natural$.



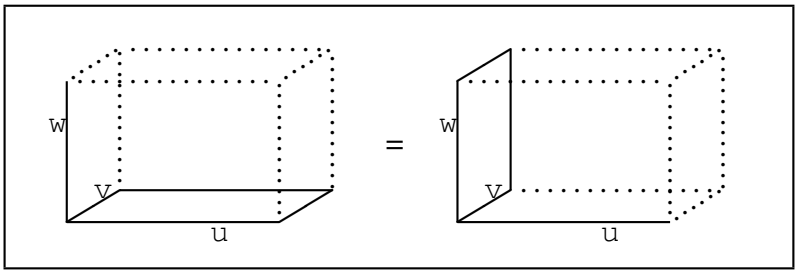
The direct-product operator $\times$ is taken to be (right-)distributive over $\sqcup$

$$\text{Symmetry 3}: \qquad (\mathtt{x} \times \mathtt{t}) \sqcup (\mathtt{y} \times \mathtt{t}) = (\mathtt{x} \sqcup \mathtt{y}) \times \mathtt{t} \tag{9}$$

so that relationships in one set, such as perhaps $\clubsuit \sqcup \heartsuit$, remain intact whether or not an independent element from the other, such as perhaps $\natural$, is appended. Left distributivity may well hold but is not needed. The direct product of independent lattices is also taken to be associative (Figure 3).

$$\text{Symmetry 4}: \qquad (\mathtt{u} \times \mathtt{v}) \times \mathtt{w} = \mathtt{u} \times (\mathtt{v} \times \mathtt{w}) \tag{10}$$

Figure 3. Associativity of direct product can be viewed geometrically.



Finally, we consider a totally ordered set of logical statements that form a chain $\mathtt{x} < \mathtt{y} < \mathtt{z} < \mathtt{t}$. We focus on an interval on the chain, which is defined by an ordered pair of logical statements $[\mathtt{x}, \mathtt{t}]$. Adjacent intervals can be chained, as in $\big[[\mathtt{x}, \mathtt{y}], [\mathtt{y}, \mathtt{z}]\big] = [\mathtt{x}, \mathtt{z}]$, and chaining is associative

$$\Big[[\mathtt{x}, \mathtt{y}], [\mathtt{y}, \mathtt{z}]\Big], [\mathtt{z}, \mathtt{t}] = [\mathtt{x}, \mathtt{y}], \Big[[\mathtt{y}, \mathtt{z}], [\mathtt{z}, \mathtt{t}]\Big] \tag{11}$$

Using Greek symbols to represent an interval, $\alpha = [\mathtt{x}, \mathtt{y}]$, $\beta = [\mathtt{y}, \mathtt{z}]$, $\gamma = [\mathtt{z}, \mathtt{t}]$, we have

$$\text{Symmetry 5 :} \qquad (\alpha, \beta), \gamma = \alpha, (\beta, \gamma) \tag{12}$$

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.

4. Axioms

We now introduce a layer of quantification. 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 to be assigned to elements of a lattice, while conforming to the above symmetries (#0—#5)for disjoint elements.

Fidelity (symmetry #0) requires us to choose an increasing measure so that, without loss of generality, we may set $m(\perp) = 0$ and thereafter

$$\text{Axiom 0 :} \qquad x > 0 \tag{13}$$

To conform to the ordering symmetry #1, we require $\oplus$ as set up in Equation 2 to obey

$$\text{Axiom 1 :} \qquad x < y \quad \Longrightarrow \quad \begin{cases} x \oplus z < y \oplus z \\ z \oplus x < z \oplus y \end{cases} \tag{14}$$

To conform to the associative symmetry #2, we also require $\oplus$ to obey

$$\text{Axiom 2 :} \qquad (x \oplus y) \oplus z = x \oplus (y \oplus z) \tag{15}$$

These equations are to hold for arbitrary values x, y, z assigned to the disjoint $\mathtt{x}$, $\mathtt{y}$, $\mathtt{z}$. Appendix A will show that these order and associativity axioms are necessary and sufficient to determine the additive calculus of measure.

To conform to the distributive symmetry #3, we require $\otimes$ as set up in Equation 3 to obey

$$\text{Axiom 3 :} \qquad (x \otimes t) \oplus (y \otimes t) = (x \oplus y) \otimes t \tag{16}$$

for disjoint $\mathtt{x}$ and $\mathtt{y}$ combined with any $\mathtt{t}$ from the second lattice. Presence of $\mathtt{t}$ may change the measures, but does not change their underlying additivity. To conform to the associative symmetry #4, we also require $\otimes$ to obey

$$\text{Axiom 4 :} \qquad (u \otimes v) \otimes w = u \otimes (v \otimes w) \tag{17}$$

These axioms determine the multiplicative form of $\otimes$ and also lead to a unique divergence between measures.

To conform to the associative symmetry #5, we require $\odot$ as set up in Equation 4 to obey

$$\text{Axiom 5 :} \qquad \Big(p(\alpha) \odot p(\beta)\Big) \odot p(\gamma) = p(\alpha) \odot \Big(p(\beta) \odot p(\gamma)\Big) \tag{18}$$

where $\alpha = [\mathtt{x}, \mathtt{y}]$, $\beta = [\mathtt{y}, \mathtt{z}]$, $\gamma = [\mathtt{z}, \mathtt{t}]$ are individual steps concatenated along the chain α, β, γ, which is $[[\mathtt{x}, \mathtt{y}], [\mathtt{y}, \mathtt{z}], [\mathtt{z}, \mathtt{t}]] = [\mathtt{x}, \mathtt{t}]$. This final axiom will let us pass from measure to probability and Bayes' theorem, and from divergence to information and entropy. For each operator (Table 1), the eventual form satisfies all relevant axioms, which assures existence. Uniqueness remains to be demonstrated.

5. Measure

Preliminary to investigating probability, we attend to the foundation of measure.

5.1. Disjoint arguments

According to the scalar *associativity theorem* (Appendix A), an operator $\oplus$ obeying axioms 1 and 2 exists and can without loss of generality be taken to be addition $+$, giving the sum rule.

$$\text{Sum rule :} \qquad x \oplus y = x + y \tag{19}$$

Commutativity $x \oplus y = y \oplus x$, though not explicitly assumed, is an unsurprising property. In accordance with fidelity (axiom 0), element values are strictly positive $x > 0$. In this form, positive-valued valuation $m(\mathtt{x}) = x$ of lattice elements is known as a *measure*. If the null element is included as the bottom of the lattice, it has zero value.

Whilst we are free to adopt additivity as a convenient convention, we are also free to adopt any order-preserving regrade Θ for which the rule would be

$$x \oplus y = \Theta^{-1}\Big(\,\Theta(x) + \Theta(y)\,\Big) \tag{20}$$

This carries no extra generality because this form 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 $\Theta(x) = Kx$, with $K > 0$ to retain positivity.

Measure theory (see for example [7]) 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 symmetries of combination require it. Any other formulation would break these basic properties of associativity and order, and would not yield a widely useful theory.

5.2. Arbitrary Arguments

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

By putting $\mathtt{x} = \mathtt{u} \sqcup \mathtt{v}$ and $\mathtt{y} = \mathtt{v} \sqcup \mathtt{w}$ for disjoint $\mathtt{u}, \mathtt{v}, \mathtt{w}$, we reach the general "inclusion/exclusion" sum rule for arbitrary $\mathtt{x}$ and $\mathtt{y}$

$$\boxed{m(\mathtt{x} \vee \mathtt{y}) + m(\mathtt{x} \wedge \mathtt{y}) = m(\mathtt{x}) + m(\mathtt{y})} \tag{21}$$

Commutativity of join and meet follow:

$$m(\mathtt{x} \vee \mathtt{y}) = m(\mathtt{y} \vee \mathtt{x})\,, \quad m(\mathtt{x} \wedge \mathtt{y}) = m(\mathtt{y} \wedge \mathtt{x})\,. \tag{22}$$

5.3. Independence

From the associativity of direct product (axiom 4), the associativity theorem (Appendix A again) assures the existence of an additivity relationship of the form

$$\Theta(x \otimes t) = \Theta(x) + \Theta(t) \tag{23}$$

for some invertible function Θ of the measures $x = m(\mathtt{x})$, $t = m(\mathtt{t})$ and $x \otimes t = m(\mathtt{x} \times \mathtt{t})$. We can not proceed as before to re-grade in terms of $\Theta(m)$ to supersede m, because we are already using additivity

$$x \otimes t + y \otimes t = (x + y) \otimes t \tag{24}$$

(axiom 3, distributivity of $\otimes$ over $\oplus$=+) to define the grade. Instead, we require consistency with the sum-rule behavior for $x \otimes t$ and $y \otimes t$. Defining $\Psi = \Theta^{-1}$ gives, term by term,

$$\Psi(\xi + \tau) + \Psi(\eta + \tau) = \Psi(\zeta(\xi, \eta) + \tau) \tag{25}$$

where

$$\xi = \Theta(x)\,, \quad \eta = \Theta(y)\,, \quad \zeta = \Theta(x+y)\,, \quad \tau = \Theta(t)\,. \tag{26}$$

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*. By definition, Ψ returns a measure, so it is positive.

The product theorem (Appendix B) shows Θ to be logarithmic, with Equation 23 reading

$$\frac{1}{A}\log\frac{x\otimes t}{C} = \frac{1}{A}\log\frac{x}{C} + \frac{1}{A}\log\frac{t}{C} \tag{27}$$

with A and C universal constants (A cancelling out), and C being positive. The obvious convention $C = 1$ loses no generality, and shows $\otimes$ to be simple multiplication

$$\text{Direct-product rule}: \qquad x\otimes t = x\,t \tag{28}$$

Measures are required to multiply, because of associativity of direct product, and the "$\otimes\, t$" operation is represented by "scale by t". This is consistent with linear rescaling (here depending on the second factor t) being the only allowed freedom for the measure assigned to the first factor x.

6. Variation

Variational principles are common in science—minimum energy for equilibrium, Hamilton's principle for dynamics, maximum entropy for thermodynamics, and so on—and we seek one for measures. The aim is to discover a variational potential $H(\mathbf{m})$ whose constrained minimum allows the valuations $\mathbf{m} = (m_1, m_2, \ldots, m_N)$ of N atoms to be assigned subject to appropriate constraints of the form $f(\mathbf{m}) =$ constant. (The vectors which appear in this section are shown in **bold-face** font.)

The variational potential is required to be general, applying to arbitrary constraints. Just like values themselves, constraints on individual atom values can be combined into compound constraints that influence several values: indeed the constraints could simply be imposition of definitive values. Such combination allows a Boolean lattice, entirely analogous to Figure 1, to be developed from individual atomic constraints. The variational potential H is to be a valuation on the measures resulting from these constraints, combination being represented by some operator $\bigcirc$ so that

$$H(x\ \texttt{WITH}\ y) = H(x)\bigcirc H(y) \tag{29}$$

for constraints acting on disjoint atoms or compounds.

Adding extra constraints always increases H, otherwise the variational requirement would be broken, so H must be faithful to chaining in the lattice.

$$\underbrace{x\ <\ y}_{\text{chained}} \qquad \Longrightarrow \qquad \underbrace{H(x)\ <\ H(y)}_{\text{real numbers}} \tag{30}$$

We also have order

$$H(x) < H(y) \quad \Longrightarrow \quad \begin{cases} H(x)\bigcirc H(z) < H(y)\bigcirc H(z) \\ H(z)\bigcirc H(x) < H(z)\bigcirc H(y) \end{cases} \tag{31}$$

because if y is a "harder" constraint than x (meaning $H(y) > H(x)$), that ranking should not be affected by some other constraint on something else. Associativity

$$\Big(H(x) \bigcirc H(y)\Big) \bigcirc H(z) = H(x) \bigcirc \Big(H(y) \bigcirc H(z)\Big) \tag{32}$$

is likewise required and expresses the combination of three constraints. It would also be natural to assume commutativity, $H(x) \bigcirc H(y) = H(y) \bigcirc H(x)$, but that is not necessary because we already recognize Equations 30–32 as our axioms 0, 1, 2. Hence, using Appendix A again, there exists a "$\bigcirc = +$" grade on which H is additive.

$$H(\mathbf{m}) = \sum_{\text{atoms } i} H_i(m_i) \tag{33}$$

We have now justified additivity, thus filling a gap in traditional accounts of the calculus of variations.

Under perturbation, the minimization requirement is

$$\delta H(\mathbf{m}) \geq 0 \quad \text{when} \quad \delta f_1(\mathbf{m}) = \delta f_2(\mathbf{m}) = \cdots = 0 \tag{34}$$

The standard "$\oplus = +$" form of the sum rule happens to be continuous and differentiable, so is applicable to valuation of systems that differ arbitrarily little. We adopt it, and can then justifiably require the variational potential to be valid for arbitrarily small perturbations:

$$dH(\mathbf{m}) = 0 \quad \text{when} \quad df_1(\mathbf{m}) = df_2(\mathbf{m}) = \cdots = 0 \tag{35}$$

This limit Equation 35 is weaker than the original Equation 34 not only because of the restricted context, but also because the nature of the extremum (maximum or minimum or saddle) is lost in the discarded second-order effects. However, it still needs to be satisfied. It also shows that any variational potential must by its nature be differentiable at least once.

One now invents supposedly constant "Lagrange multiplier" coefficients $\lambda_1, \lambda_2, \ldots$ and considers what appears at first to be the different problem of solving

$$d\Big(\, H(\mathbf{m}) - \lambda_1 f_1(\mathbf{m}) - \lambda_2 f_2(\mathbf{m}) - \ldots \Big) = 0 \quad \text{under arbitrary perturbation} \tag{36}$$

for $\mathbf{m}$. Clearly, Equation 36 is equivalent to Equation 35 for perturbations that happen to hold the f's constant ($df = 0$). However, the values those f's take may well be wrong. The trick is to choose the λ's so that the f's take their correct constraint values. That being done, Equation 36 solves the variational problem Equation 35.

Let the application be two-dimensional, x-by-y, in the sense of applying to values $m(\mathrm{x} \times \mathrm{y})$ of elements on a direct-product lattice. Suppose we have x-dependent constraints that yield $m(\mathrm{x}) = m_x$ on one factor (say the card suits in Figure 2 above), and similar y-dependent constraints that yield $m(\mathrm{y}) = m_y$ on the other factor (say music keys in Figure 2). Both factors being thus controlled, their direct-product is implicitly controlled by the those same constraints. Here, we already know the target value $m(\mathrm{x} \times \mathrm{y}) = m_x m_y$ from the direct-product rule Equation 28. Hence the variational assignment for the particular value $m(\mathrm{x} \times \mathrm{y})$ derives from

$$H'_{xy}(m_x m_y) = \lambda_1 f_1(m_x) + \lambda_2 f_2(m_y) \tag{37}$$

(where $'$ indicates derivative). The variational theorem (Appendix C) gives the solution of this functional equation as

$$H_i(m_i) = A_i + B_i m_i + C_i(m_i \log m_i - m_i) \tag{38}$$

for the individual valuation being considered, where A_i, B_i, C_i are constants. Combining all the atoms yields

$$\boxed{H(\mathbf{m}) = \sum_{\text{atoms } i} \Big(A_i + B_i m_i + C_i(m_i \log m_i - m_i)\Big)} \tag{39}$$

The coefficient C_i represents the intrinsic importance of atom $\mathtt{a}_i$ in the summation, but usually the atoms are *a priori* equivalent so that the C's take a common value. The scaling of a variational potential is arbitrary (and is absorbed in the Lagrange multipliers), 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.

6.1. Divergence and Distance

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 that led to $\mathbf{w}$ 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

$$\text{Divergence}: \qquad H(\mathbf{w} \mid \mathbf{u}) = \sum_{\text{atoms } i} \Big(u_i - w_i + w_i \log(w_i/u_i)\Big) \tag{40}$$

This formula is unique: none other has the properties Equations 33,37 that elementary applications require. Equivalently, any different formula would give unjustifiable answers in those applications. Plausibly, H is non-negative, $H(\mathbf{w} \mid \mathbf{u}) \geq 0$ with equality if and only if $\mathbf{w} = \mathbf{u}$, so that it usefully quantifies the separation of destination from source.

In general, H obeys neither commutativity nor the triangle inequality, $H(\mathbf{w} \mid \mathbf{u}) \neq H(\mathbf{u} \mid \mathbf{w})$ and $H(\mathbf{w} \mid \mathbf{u}) \not\leq H(\mathbf{w} \mid \mathbf{v}) + H(\mathbf{v} \mid \mathbf{u})$. Hence it cannot be a geometrical "distance", which is required to have both those properties. In fact, there is no definition of geometrical measure-to-measure distance that obeys the basic symmetries, because H is the only candidate, and it fails.

Here again we see our methodology yielding clear insight. "From–to" can be usefully quantified, but "between" cannot. A space of measures may have connectedness, continuity, even differentiability, but it cannot become a metric space and remain consistent with its foundation.

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

$$H(\mathbf{w} \mid \mathbf{u}) = \int \Big(u(\theta) - w(\theta) + w(\theta) \log(w(\theta)/u(\theta))\Big)\, d\theta \tag{41}$$

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. This is the rationale behind maximum entropy data analysis [8].

7. Probability Calculus

In inference, we seek to impose on the hypothesis space a quantified *degree of implication* $p(\mathtt{x} \mid \mathtt{t})$, to represent the plausibility of predicate $\mathtt{x}$ conditional on current knowledge that excludes all hypotheses outside the stated context $\mathtt{t}$. This is accomplished via a bivaluation, which is a functional that takes a pair of lattice elements to a real number. This bivaluation should depend on both $\mathtt{x}$ (obviously) and $\mathtt{t}$ (otherwise it would be just the measure assigned to $\mathtt{x}$). The natural conjecture is that probability should be identified with a normalized measure, and we proceed to prove this—measures can have arbitrary total but probabilities will (according to standard convention) sum to unity.

At the outset, though, we simply wish to set up a bivaluation for predicate $\mathtt{x}$ within context $\mathtt{t}$.

7.1. Chained Arguments

Within given context $\mathtt{t}$, we require $p(\mathtt{x} \mid \mathtt{t})$ to have the order and associative symmetries #1 and #2 that define a measure. Consequently, p obeys the sum rule

$$p(\mathtt{x} \sqcup \mathtt{y} \mid \mathtt{t}) = p(\mathtt{x} \mid \mathtt{t}) + p(\mathtt{y} \mid \mathtt{t}) \tag{42}$$

for disjoint $\mathtt{x}$ and $\mathtt{y}$ with $\mathtt{x} \sqcup \mathtt{y} < \mathtt{t}$. It is the dependence on $\mathtt{t}$ that remains to be determined.

Associativity of chaining (axiom 5) for $\mathtt{a} < \mathtt{b} < \mathtt{c} < \mathtt{d}$ is represented by

$$\Big(\underbrace{\underbrace{p(\mathtt{a} \mid \mathtt{b})}_{p(\alpha)} \odot \underbrace{p(\mathtt{b} \mid \mathtt{c})}_{p(\beta)}}_{p(\alpha,\beta)}\Big) \odot \underbrace{p(\mathtt{c} \mid \mathtt{d})}_{p(\gamma)} = \underbrace{p(\mathtt{a} \mid \mathtt{b})}_{p(\alpha)} \odot \Big(\underbrace{\underbrace{p(\mathtt{b} \mid \mathtt{c})}_{p(\beta)} \odot \underbrace{p(\mathtt{c} \mid \mathtt{d})}_{p(\gamma)}}_{p(\beta,\gamma)}\Big) \tag{43}$$

We do not have commutativity, $(\alpha, \beta) = [[\mathtt{a}, \mathtt{b}], [\mathtt{b}, \mathtt{c}]] = [\mathtt{a}, \mathtt{c}]$ not being the same as (β, α) (which is meaningless), but we do have associativity and we do have order along the chain. By the associativity theorem, $\odot$ exists and there is a scale on which it is simple addition. However, we can not regrade to that scale and discard the original because we have already fixed the grade of p to be additive with respect to its first argument. Instead, we infer additivity on some other grade $\Theta(p)$

$$\Theta\Big(\underbrace{p(\mathtt{a} \mid \mathtt{c})}_{p(\alpha)\odot p(\beta)}\Big) = \Theta\Big(\underbrace{p(\mathtt{a} \mid \mathtt{b})}_{p(\alpha)}\Big) + \Theta\Big(\underbrace{p(\mathtt{b} \mid \mathtt{c})}_{p(\beta)}\Big) \tag{44}$$

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

$$\underbrace{p(\mathtt{a} \mid \mathtt{c})}_{p(\alpha)\odot p(\beta)} = \Psi\Big(\Theta(\underbrace{p(\mathtt{a} \mid \mathtt{b})}_{p(\alpha)}) + \Theta(\underbrace{p(\mathtt{b} \mid \mathtt{c})}_{p(\beta)})\Big) \tag{45}$$

Substituting this in the sum rule Equation 42, term by term, yields the same "product Equation" 25

$$\Psi(\zeta(\xi, \eta) + \tau) = \Psi(\xi + \tau) + \Psi(\eta + \tau) \tag{46}$$

as before, where

$$\xi = \Theta\Big(p(\mathtt{x} \mid \mathtt{z})\Big)\,, \quad \eta = \Theta\Big(p(\mathtt{y} \mid \mathtt{z})\Big)\,, \quad \zeta = \Theta\Big(p(\mathtt{x} \sqcup \mathtt{y} \mid \mathtt{z})\Big)\,, \quad \tau = \Theta\Big(p(\mathtt{z} \mid \mathtt{t})\Big)\,. \tag{47}$$

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

$$p(\mathtt{x} \mid \mathtt{z}) = p(\mathtt{x} \mid \mathtt{y})\, p(\mathtt{y} \mid \mathtt{z}) \,/\, C \tag{48}$$

in which p (positive by virtue of being a measure on predicates) 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.

$$\text{Chain-product rule :} \qquad p(\mathtt{x} \mid \mathtt{z}) = p(\mathtt{x} \mid \mathtt{y})\, p(\mathtt{y} \mid \mathtt{z}) \tag{49}$$

7.2. Arbitrary Arguments

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 $\mathtt{x} \wedge \mathtt{y} = \mathtt{x}$ in a chain where $\mathtt{x} < \mathtt{y}$, so that $p(\mathtt{x} \wedge \mathtt{y} \mid \mathtt{y}) = p(\mathtt{x} \mid \mathtt{y})$. The general form

$$p(\mathtt{a} \wedge \mathtt{b} \mid \mathtt{c}) = p(\mathtt{a} \mid \mathtt{b} \wedge \mathtt{c})\, p(\mathtt{b} \mid \mathtt{c}) \tag{50}$$

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

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

Probability calculus is now proved:

Range	$0 = \mathrm{Pr}(\bot \mid \mathtt{t}) < \mathrm{Pr}(\mathtt{x} \mid \mathtt{t}) \leq \mathrm{Pr}(\mathtt{t} \mid \mathtt{t}) = 1$
Sum rule	$\mathrm{Pr}(\mathtt{x} \vee \mathtt{y} \mid \mathtt{t}) + \mathrm{Pr}(\mathtt{x} \wedge \mathtt{y} \mid \mathtt{t}) = \mathrm{Pr}(\mathtt{x} \mid \mathtt{t}) + \mathrm{Pr}(\mathtt{y} \mid \mathtt{t})$
Chain-product	$\mathrm{Pr}(\mathtt{x} \wedge \mathtt{y} \mid \mathtt{t}) = \mathrm{Pr}(\mathtt{x} \mid \mathtt{y} \wedge \mathtt{t})\ \mathrm{Pr}(\mathtt{y} \mid \mathtt{t})$

The top element of the current lattice, $\mathtt{t}$, is the (provisional) truth, often written $\top$.

From the commutativity $\mathrm{Pr}(\mathtt{x} \wedge \mathtt{y} \mid \mathtt{t}) = \mathrm{Pr}(\mathtt{y} \wedge \mathtt{x} \mid \mathtt{t})$ associated with $\wedge$, we obtain Bayes' Theorem

$$\mathrm{Pr}(\mathtt{x} \mid \theta \wedge \mathtt{t})\, \mathrm{Pr}(\theta \mid \mathtt{t}) = \mathrm{Pr}(\theta \mid \mathtt{x} \wedge \mathtt{t})\, \mathrm{Pr}(\mathtt{x} \mid \mathtt{t}) \tag{51}$$

which can be simplified by making the common context implicit and writing

$$\underbrace{\mathrm{Pr}(\mathtt{x} \mid \theta)}_{\text{Likelihood}}\ \underbrace{\mathrm{Pr}(\theta)}_{\text{Prior}} = \underbrace{\mathrm{Pr}(\theta \mid \mathtt{x})}_{\text{Posterior}}\ \underbrace{\mathrm{Pr}(\mathtt{x})}_{\text{Evidence}} \qquad \| \ \mathtt{t} \tag{52}$$

to relate data $\mathtt{x}$ and parameter θ (with context $\mathtt{t}$ understood). Do not misinterpret the abbreviated notation. Probability is always and necessarily, by construction, a bivaluation that assigns a real number to a *pair* of elements in a Boolean lattice. In addition, one does not need to differentiate between

likelihood, prior, posterior, and evidence by giving each one a different notation. The terms that comprise Bayes' Theorem represent the same bivaluation applied to different pairs of elements.

7.3. Probability as a Ratio

The equations of probability calculus (range, sum rule, and chain-product rule) can all be subsumed in the single expression

$$\Pr(\mathtt{x} \mid \mathtt{t}) = \frac{m(\mathtt{x} \wedge \mathtt{t})}{m(\mathtt{t})} \qquad \forall \mathtt{x},\ \forall \mathtt{t} \neq \bot \tag{53}$$

for probability as a ratio of measures. Thus the calculus of probability is nothing more than the elementary calculus of proportions of measure. As anticipated, within its context $\mathtt{t}$, 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 [9]) 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 measures in $[0, \infty)$, so ordinary proportions in $[0, 1]$ are forced for probability calculus.

Whereas the sum rule for measure and probability generalizes to the inclusion/exclusion form for general elements which need not be disjoint, so does the ratio form of probability allow generalization from intervals [3] to ***generalized intervals***, consisting of arbitrary pairs $[\mathtt{x}, \mathtt{t}]$ which need not be in a chain. The bivaluation form Equation 53 still holds but now represents a general ***degree of implication*** between arbitrary elements.

8. Information and Entropy

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

8.1. 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

$$\text{Information}: \qquad \boxed{H(\mathbf{p} \mid \mathbf{q}) = \sum_k p_k \log \frac{p_k}{q_k}} \tag{54}$$

In statistics, this is known as the Kullback–Leibler formula [10].

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

$$H(\mathbf{p} \mid \mathbf{q}) = -\log q_k \quad \text{when } p_k = 1\,. \tag{55}$$

This represents the information gained on acquiring the knowledge that the specific k was true—equivalently the surprise at finding k instead of any available alternative. Generally, H is the amount of *compression* (logarithmically, with respect to the source) induced by the constraints that modulate source into destination.

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

$$H(\mathbf{p} \mid \mathbf{q}) = \int p(x) \log \frac{p(x)}{q(x)}\, dx \tag{56}$$

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

8.2. Entropy

The variational potential

$$H(\mathbf{p}) = \sum_k \Big(A_k + B_k p_k + C(p_k \log p_k - p_k) \Big) \tag{57}$$

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

$$\texttt{Entropy}: \qquad \boxed{S(\mathbf{p}) = -\sum_k p_k \log p_k} \tag{58}$$

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.

$$S(\mathbf{p}) = \big\langle -\log p_k \big\rangle_k \tag{59}$$

It is a function of the partitioning as well as the probability distribution, which is why it does not 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 [11] 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. Its value generally diverges upwards as the partitioning deepens, whereas H usually tends towards a continuum limit.

9. Conclusions

9.1. Summary

We start with a set $\{\mathtt{a}_1, \mathtt{a}_2, \mathtt{a}_3, \ldots, \mathtt{a}_N\}$ of "atomic" elements which in inference represent the most fundamental exclusive statements we can make about the states (of our model) of the world. Atoms combine to form a Boolean lattice which in inference is called the hypothesis space of statements. This structure has rich symmetry, but other applications may have less and we have selected only what we needed, so that our results apply more widely and to distributive lattices in particular. The minimal assumptions are so simple that they can be drawn as the cartoon below (Figure 4).

Axiom 1 represents the order property that is required of the combination operator $\sqcup$. Axiom 2 says that valuation must conform to the associativity of $\sqcup$. These axioms are compelling 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 $m(\mathtt{x})$, with $\sqcup$ 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 3) and associative (axiom 4), 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 $\mathtt{x}$ within any specified context $\mathtt{t}$. Axiom 5 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.

9.2. Commentary

We have presented a foundation for inference that unites and significantly extends the approaches of Kolmogorov [2] and Cox [1], 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 minimal required 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 negation (`NOT`), which in applications other than inference may well not be present. Neither have we assumed any additive or multiplicative behavior (as did Kolmogorov [2], de Finetti [12], and Dupré & Tipler [13]). On the contrary, we find that sum and product rules follow from elementary symmetry alone.

Figure 4. Cartoon graphic of the symmetries invoked, and where they lead. Ordering is drawn as upward arrows.

Fidelity		
0: Values increase	$\begin{array}{c}\spadesuit\\ \uparrow\\ \clubsuit\end{array}$	**Positivity**
Combination		
1: $\sqcup$ preserves order	$\begin{array}{c}\spadesuit\\ \uparrow\\ \clubsuit\end{array} \Longrightarrow \begin{array}{c}\spadesuit\sqcup\heartsuit\\ \uparrow\\ \clubsuit\sqcup\heartsuit\end{array}$ and $\begin{array}{c}\heartsuit\sqcup\spadesuit\\ \uparrow\\ \heartsuit\sqcup\clubsuit\end{array}$	
2: $\sqcup$ is associative	$\underbrace{(\underbrace{\clubsuit\sqcup\heartsuit})\sqcup\spadesuit} = \underbrace{\clubsuit\sqcup(\underbrace{\heartsuit\sqcup\spadesuit})}$	**Measure**
Direct product		
3: $\times$ is distributive	$\left.\begin{array}{c}\clubsuit\sqcup\heartsuit\\ \nearrow\quad\nwarrow\\ \clubsuit\qquad\heartsuit\end{array}\right\}\times\natural = \begin{array}{c}(\clubsuit\sqcup\heartsuit)\times\natural\\ \nearrow\quad\nwarrow\\ \clubsuit\times\natural\quad\heartsuit\times\natural\end{array}$	
4: $\times$ is associative	$\underbrace{(\underbrace{\clubsuit\times\natural})\times\aleph} = \underbrace{\clubsuit\times(\underbrace{\natural\times\aleph})}$	**Divergence**
Implication		
5: chaining is associative	$\begin{array}{c}\diamondsuit\\ \uparrow\gamma\\ \heartsuit\\ \uparrow\beta\\ \spadesuit\\ \uparrow\alpha\\ \clubsuit\end{array} = \begin{array}{c}\diamondsuit\\ \uparrow\gamma\\ \heartsuit\\ \uparrow\beta\\ \spadesuit\\ \uparrow\alpha\\ \clubsuit\end{array}$	**Measure** $\longrightarrow$ **Probability and Bayes** **Divergence** $\longrightarrow$ **Information and Entropy**

We find that associativity and order provide minimal assumptions that are convincing and compelling for scalar additivity in all its applications. Associativity alone does not force additivity, but associativity with order does. Positivity was not assumed, though it holds for all applications in this paper.

Commutativity was not assumed either, though commutativity of the resulting measure follows as a property of additivity. Associativity and commutativity do not quite force additivity because they allow degenerate solutions such as $a \oplus b = \max(a, b)$. To eliminate these, strict order is required in some form, and if order is assumed then commutativity does not need to be. Hence scalar additivity rests on ordered sequences rather than the disordered sets for which commutativity would be axiomatic.

$$\begin{array}{lcl} \text{Associativity} + \text{Order} & \Longrightarrow & \text{Additivity allowed} \Longrightarrow \text{Commutativity} \\ \text{Associativity alone} & \not\Longrightarrow & \text{Additivity allowed} \\ \text{Associativity} + \text{Commutativity} & \not\Longrightarrow & \text{Additivity allowed} \end{array}$$

Aczél [14] assumes order in the form of reducibility, and he too derives commutativity. However, his analysis assumes the continuum limit already attained, which requires him to assume continuity.

$$\text{Associativity} + \text{Order} + \text{Continuity} \Longrightarrow \text{Additivity allowed} \Longrightarrow \text{Commutativity}$$

Our constructivist approach uses a finite environment in which continuity does not apply, and proceeds directly to additivity. Here, continuity and differentiability are merely emergent properties of $+$ as the continuum limit is approached by allowing arbitrarily many atoms of different type.

Yet there can be no *requirement* of continuity, which is merely a convenient *convention*. For example, re-grading could take the binary representations of standard arguments (101.011_2 representing $5\frac{3}{8}$) and interpret them in base-3 ternary (with 101.011_3 representing $10\frac{4}{27}$), so that $\Theta(10\frac{4}{27}) = 5\frac{3}{8}$. Valuation becomes discontinuous everywhere, but the sum rule still works, albeit less conveniently. Indeed, no finite system can ever demonstrate the infinitesimal discrimination that defines continuity, so continuity cannot possibly be a requirement of practical inference.

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 over infinite sets and the rôle of the continuum disappears.

Other than reversible regrading, any deviation from the standard formulas must inevitably contradict the elementary symmetries that underlie them, so that popular but weaker justifications (e.g., de Finetti [12]) in terms of decisions, loss functions, or monetary exchange can be discarded as unnecessary. In fact, the logic is the other way round: such applications must be cast in terms of the unique calculus of measure and probability if they are to be quantified rationally. 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 an illogical impression that the strong argument might have been circumvented too.

Finally, the approach from basic symmetry is productive. Goyal and ourselves [15] 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 Seth Chaikin, Janos Aczél, Ariel Caticha, Julian Center, Philip Goyal, Steve Gull, Jeffrey Jewell, Vassilis Kaburlasos, Carlos Rodríguez, and a thoughtful anonymous reviewer. 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.

References

1. Cox, R.T. Probability, frequency, and reasonable expectation. *Am. J. Phys.* **1946**, *14*, 1–13.
2. Kolmogorov, A.N. *Foundations of the Theory of Probability*, 2nd English ed.; Chelsea: New York, NY, USA, 1956.
3. Birkhoff, G. *Lattice Theory*; American Mathematical Society: Providence, RI, USA, 1967.
4. Davey, B.A.; Priestley, H.A. *Introduction to Lattices and Order*; Cambridge University Press: Cambridge, UK, 2002.
5. Klain, D.A.; Rota, G.-C. *Introduction to Geometric Probability*; Cambridge University Press: Cambridge, UK, 1997.
6. Knuth, K.H. Deriving Laws from Ordering Relations. In *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*; Erickson, G.J., Zhai, Y., Eds.; Jackson: Hole, WY, USA, 2003.
7. Halmos, P.R. *Measure Theory*; Springer: Berlin/Heidelberg, Germany; 1974.
8. Gull, S.F.; Skilling, J. Maximum entropy method in image processing. *IEE Proc. 131F*, 646–659.
9. Von Mises, R. *Probability, Statistics, and Truth*; Dover: Mineola, NY, USA, 1981.
10. Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann. Math. Statist.* **1951**, *22*, 79–86.
11. Shannon, C.F. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423, 623–656.
12. De Finetti, B. *Theory of Probability, Vol. I and Vol. II*; John Wiley and Sons: New York, NY, USA, 1974.
13. Dupré, M.J.; Tipler, F.J. New axioms for rigorous Bayesian probability. *Bayesian Anal.* **2009**, *4*, 599–606.
14. Aczél, J. *Lectures on Functional Equations and Their Applications*; Academic Press: New York, NY, USA, 1966.
15. Goyal, P.; Knuth, K.H.; Skilling, J. Origin of complex quantum amplitudes and Feynman's rules. *Phys. Rev. A* **2010**, *81*, 022109.

A. Appendix A: Associativity Theorem

Atoms $\mathtt{x}, \mathtt{y}, \mathtt{z}, \ldots$, or disjoint lattice elements more generally, are to be assigned valuations $x, y, z, \ldots$. If valuations coincide (though other marks may differ), such atoms are said to be of the same type. We allow arbitrarily many atoms of arbitrarily many types. Our proof is constructive, with combinations

built as sequences of atoms appended one at a time, $\mathtt{x} \sqcup \mathtt{y} \sqcup \ldots$ having valuation $x \oplus y \oplus \ldots$. The consequent stand-alone derivation is rather long, but avoids making what would in our finite environment be an unnatural assumption of continuity. We also avoid assuming that an inverse to combination exists.

We merely assume order (axiom 1)

$$\begin{array}{ll} \text{Axiom 1a :} & \\ \text{Axiom 1b :} & \end{array} \qquad x < y \quad \Longrightarrow \quad \left\{ \begin{array}{l} x \oplus z < y \oplus z \\ z \oplus x < z \oplus y \end{array} \right.$$

and associativity (axiom 2)

$$\text{Axiom 2 :} \qquad (x \oplus y) \oplus z = x \oplus (y \oplus z)$$

Theorem:

> Axiom 1 (order) and axiom 2 (associativity) imply that
>
> $$x \oplus y = \Theta^{-1}\Big(\Theta(x) + \Theta(y)\Big)$$
>
> for any order-preserving regrade Θ of "$\oplus = +$" applied to scalar values.

A.1. **Proof:**

The form quoted in the theorem is easily seen to satisfy both axioms 1 and 2, which demonstrates *existence* of a calculus $\oplus$ of quantification. The remaining question is whether this calculus is *unique*.

We start by building sequences from just one type of atom before introducing successively more types to reach the general case. In this way, we lay down successively finer grids. Whenever another atom is introduced to generate a new sequence, that new sequence's value inevitably lies somewhere at, between, or beyond previously assigned values. If it lies within an interval, we are free to choose it to be anywhere convenient. Such choice loses no generality, because the original value could be recovered by order-preserving regrade of the assignments. Values can be freely and reversibly regraded *in and only in* any way that preserves their order. Any such mapping preserves axiom 1, but reversal of ordering would allow the axiom to be broken.

Most points of the continuum escape this approach and are never accessed, so we do not allow ourselves continuum properties such as continuity. We build our finite system from the bottom up, using only those values that we actually need.

By interchanging x and y in axiom 1, the same relationship holds when "$<$" is replaced throughout by "$>$", and replacement by "$=$" holds trivially. So, in effect, the axiom makes a three-fold assertion

$$x \left\{ \begin{array}{c} < \\ = \\ > \end{array} \right\} y \qquad \Longrightarrow \qquad x \oplus z \left\{ \begin{array}{c} < \\ = \\ > \end{array} \right\} y \oplus z \quad \text{and} \quad z \oplus x \left\{ \begin{array}{c} < \\ = \\ > \end{array} \right\} z \oplus y \tag{60}$$

Because these three possibilities $(<, >, =)$ are exhaustive, consistency implies the reverse, sometimes called "cancellativity":

$$x \oplus z \left\{ \begin{array}{c} < \\ = \\ > \end{array} \right\} y \oplus z \quad \text{or} \quad z \oplus x \left\{ \begin{array}{c} < \\ = \\ > \end{array} \right\} z \oplus y \qquad \Longrightarrow \qquad x \left\{ \begin{array}{c} < \\ = \\ > \end{array} \right\} y \tag{61}$$

A.2. One Type of Atom

Consider a set of disjoint atoms $\{\mathtt{a}_1, \mathtt{a}_2, \mathtt{a}_3, \ldots, \mathtt{a}_r, \mathtt{a}_{r+1}, \ldots, \mathtt{a}_N\}$, each of which is associated with the same value so that $m(\mathtt{a}_i) = a$ for all $i \in [1, N]$. We will append such atoms one at a time, using the combination operator $\sqcup$ to construct compound elements

$$(((\mathtt{a}_1 \sqcup \mathtt{a}_2) \sqcup \ldots \sqcup)\mathtt{a}_r) \sqcup \mathtt{a}_{r+1} \tag{62}$$

which are to be valued as

$$(((m(\mathtt{a}_1) \oplus m(\mathtt{a}_2)) \oplus \ldots) \oplus m(\mathtt{a}_r)) \oplus m(\mathtt{a}_{r+1})\,. \tag{63}$$

Since the atoms $\mathtt{a}_i$ all have the same value, the subscripts are immaterial for valuation and we may write

$$\text{“1 of a”} \equiv \mathtt{a}_1, \quad \text{so that } m(1 \text{ of } \mathtt{a}) = m(\mathtt{a}_1) = m(\mathtt{a}) = a \tag{64}$$

and

$$\text{“2 of a”} \equiv \mathtt{a}_1 \sqcup \mathtt{a}_2, \quad \text{so that } m(2 \text{ of } \mathtt{a}) = m(\mathtt{a}_1 \sqcup \mathtt{a}_2) = m(\mathtt{a} \sqcup \mathtt{a}) \tag{65}$$

and so on with the addition of

$$\text{“0 of a”} \equiv \emptyset, \quad \text{so that } m(0 \text{ of } \mathtt{a}) = m(\emptyset) \equiv m_\emptyset\,. \tag{66}$$

In principle, we could have any of

$$m(0 \text{ of } \mathtt{a}) \left\{ \begin{array}{c} < \\ = \\ > \end{array} \right\} m(1 \text{ of } \mathtt{a}) \qquad \left\{ \begin{array}{l} \text{positive style} \\ \text{null style} \\ \text{negative style} \end{array} \right. \tag{67}$$

Null-style atoms all share the same value $m_\emptyset$. If there were two such values, say $m_\emptyset$ and $m'_\emptyset$, then the equalities

$$m(\mathtt{x}) = m_\emptyset \oplus m(\mathtt{x}) = m'_\emptyset \oplus m(\mathtt{x}) \tag{68}$$

for any $\mathtt{x}$ would, by cancellativity, make them equal.

We proceed with atoms restricted to positive style, leaving the extension to negative (if required) until the end. Chaining a sequence of positive $\mathtt{a}$'s with another $\mathtt{a}$ yields, successively, the same nature of relationship between $m(1 \text{ of } \mathtt{a})$ and $m(2 \text{ of } \mathtt{a})$, then $m(2 \text{ of } \mathtt{a})$ and $m(3 \text{ of } \mathtt{a})$, and by induction $m(r \text{ of } \mathtt{a})$ and $m(r{+}1 \text{ of } \mathtt{a})$. Hence successive multiples are ranked by cardinality, and can continue indefinitely.

$$m(\emptyset) < m(1 \text{ of } \mathtt{a}) < m(2 \text{ of } \mathtt{a}) < \cdots < m(r \text{ of } \mathtt{a}) < m(r{+}1 \text{ of } \mathtt{a}) < \ldots \tag{69}$$

Whatever values were initially proposed, we are free to regrade to other values of our choice, provided only that relevant order is preserved. Here, we are free to assign values as multiples

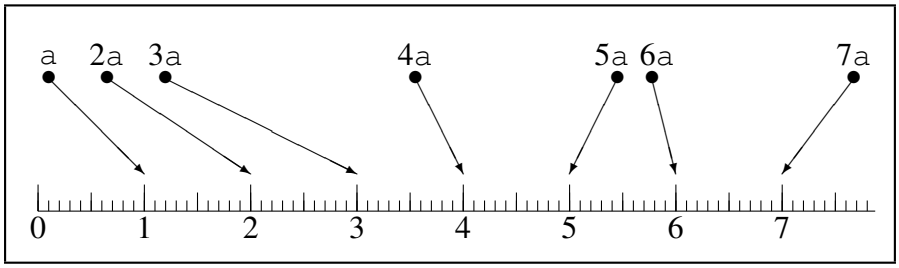
$$m(r \text{ of } \mathtt{a}) = ra \tag{70}$$

of any positive value $a > 0$. The basic linear additive scale is now in place.

Illustration

We are not forced to adopt this linear scale, and a user's original assignments may well not have used it. We can allow other increasing series, such as $m(r \text{ of } \mathtt{a}) = r^3 a$, but we could not use a non-increasing series like $m(r \text{ of } \mathtt{a}) = a \sin(r)$ without some values being the wrong way round. The only acceptable grades preserve order so that they can be monotonically reverted to the adopted integer scale (Figure 5).

Figure 5. Ordered multiples can be placed on an integer scale, here drawn with $a = 1$.



A.3. Induction to More Than One Type of Atom

Suppose that sequences of atoms drawn from up to k types $\{\mathtt{a}, \ldots, \mathtt{c}\}$ are quantified as the grid of values

$$\mu(r, \ldots, t) \equiv \underbrace{m(r \text{ of } \mathtt{a} \text{ and } \ldots \text{ and } t \text{ of } \mathtt{c})}_{\text{multiples of up to } k \text{ types in any order}} = \underbrace{ra + \ldots + tc}_{\text{corresponding terms}} \tag{71}$$

for positive multiples $r, \ldots, t$. Any individual marks that the atoms may possess beyond their type are ignored in this scalar representation. This hypothesis Equation 71 is already the assignment for $k = 1$, and we aim to develop it to all k by induction. Before doing this, we note that commutativity is implicit in Equation 71 for atoms or sequences drawn from the original k types, because

$$\mu(r + r', \ldots, t + t') = \mu(r, \ldots, t) + \mu(r', \ldots, t') \tag{72}$$

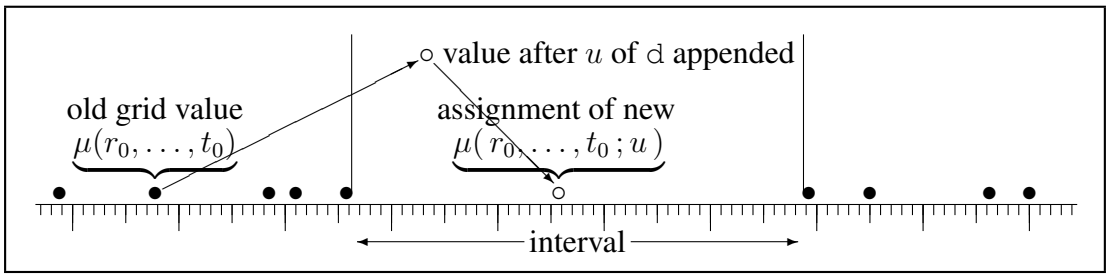
But commutativity for $k > 1$ is not being improperly assumed, because the inductive proof starts from $k = 1$, for which Equation 71 reduces to the proven Equation 70.

We now append an extra type $\mathtt{d}$ of atom, and investigate values of the extended function

$$\mu(r, \ldots, t\,; u) = m(r \text{ of } \mathtt{a} \text{ and } \ldots \text{ and } t \text{ of } \mathtt{c}) \oplus m(u \text{ of } \mathtt{d}) \tag{73}$$

formed by appending, successively, $u = 1, 2, 3, \ldots$ new atoms. If a new value coincides with an already-assigned value, it is thereby determined. Otherwise, the new value must interleave (including lying beyond) existing ones, and we are free to assign it any convenient value within that particular interval (Figure 6).

Figure 6. A new value, displaced away from the existing grid, must lie within some interval. Any assignment outside the strict interior would be wrongly ordered, while any value inside could be reverted to some other selection by order-preserving regrade.



A.3.1 *Repetition Lemma*

To proceed, we need the repetition lemma, that if

$$\mu(r, \ldots, t) \left\{ \begin{array}{c} < \\ = \\ > \end{array} \right\} \mu(r_0, \ldots, t_0\,; u) \tag{74}$$

then

$$\mu(nr, \ldots, nt) \left\{ \begin{array}{c} < \\ = \\ > \end{array} \right\} \mu(nr_0, \ldots, nt_0\,; nu) \tag{75}$$

for n-fold repetition.

Suppose the lemma does hold for n. Prefix Equation 74 with "nr_0 of a and ... and nt_0 of c", and postfix with nu of d.

$$\mu(nr_0{+}r, \ldots, nt_0{+}t) \oplus m(nu \text{ of d}) \left\{ \begin{array}{c} < \\ = \\ > \end{array} \right\} \mu\big((n{+}1)r_0, \ldots, (n{+}1)t_0\,;(n{+}1)u\big) \tag{76}$$

Prefix Equation 75 with "r of a and ... and t of c".

$$\mu\big((n{+}1)r, \ldots, (n{+}1)t\big) \left\{ \begin{array}{c} < \\ = \\ > \end{array} \right\} \mu\big(nr_0{+}r, \ldots, nt_0{+}t\,; nu\big) \tag{77}$$

Because

$$\mu(nr_0{+}r, \ldots, nt_0{+}t) \oplus m(nu \text{ of d}) = \mu\big(nr_0{+}r, \ldots, nt_0{+}t\,; nu\big) \tag{78}$$

(these two expressions being alternative notations for the same quantity), the relationships Equation 77 and Equation 76 combine to give

$$\mu\big((n{+}1)r, \ldots, (n{+}1)t\big) \left\{ \begin{array}{c} < \\ = \\ > \end{array} \right\} \mu\big((n{+}1)r_0, \ldots, (n{+}1)t_0\,;(n{+}1)u\big) \tag{79}$$

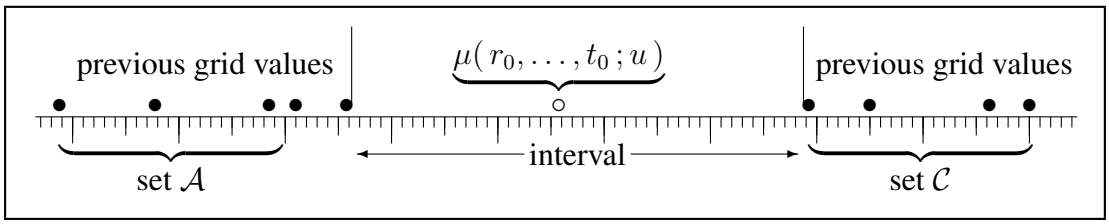
So, if Equation 75 holds for n, it also holds for $n+1$. It does hold for $n=1$, proving by induction the repetition lemma for all $n = 1, 2, 3, \ldots$.

A.3.2 *Separation*

We define the relevant intervals for the new sequences $\mu(r_0, \ldots, t_0\,; u)$ by listing the previous values Equation 71 that lie below (set $\mathcal{A}$), at (set $\mathcal{B}$), and above (set $\mathcal{C}$) the new targets (Figure 7).

$$\left\{ \begin{array}{c} \mathcal{A} \\ \mathcal{B} \\ \mathcal{C} \end{array} \right\} : \qquad \{r, \ldots, t\,; u\} \quad \text{such that} \quad \mu(r, \ldots, t) \left\{ \begin{array}{c} < \\ = \\ > \end{array} \right\} \mu(r_0, \ldots, t_0\,; u) \tag{80}$$

Figure 7. The interval encompassing the new value lies above set $\mathcal{A}$ and below set $\mathcal{C}$.



This decomposition must hold consistently across all new sequences, for all u. Values for any particular target multiplicity u lie in subsets of $\mathcal{A}, \mathcal{B}, \mathcal{C}$ with u fixed appropriately. It is convenient to denote provenance with a suffix (1 for $\mathcal{A}$, 2 for $\mathcal{B}$, 3 for $\mathcal{C}$), so that these definitions can be alternatively written as

$$\begin{array}{lll} \mathcal{A}: & \{r_1, \ldots, t_1\,; u_1\} & \text{such that} \quad \mu(r_1, \ldots, t_1) < \mu(r_0, \ldots, t_0\,; u_1) \\ \mathcal{B}: & \{r_2, \ldots, t_2\,; u_2\} & \text{such that} \quad \mu(r_2, \ldots, t_2) = \mu(r_0, \ldots, t_0\,; u_2) \\ \mathcal{C}: & \{r_3, \ldots, t_3\,; u_3\} & \text{such that} \quad \mu(r_3, \ldots, t_3) > \mu(r_0, \ldots, t_0\,; u_3) \end{array} \tag{81}$$

Apply repetitions $n = u_2 u_3$ for set $\mathcal{A}$, and $n = u_1 u_3$ for set $\mathcal{B}$, and $n = u_1 u_2$ for set $\mathcal{C}$.

$$\begin{array}{ll} \mathcal{A}: & \mu(u_2 u_3 r_1, \ldots, u_2 u_3 t_1) < \mu(u_2 u_3 r_0, \ldots, u_2 u_3 t_0\,; u_1 u_2 u_3) \\ \mathcal{B}: & \mu(u_1 u_3 r_2, \ldots, u_1 u_3 t_2) = \mu(u_1 u_3 r_0, \ldots, u_1 u_3 t_0\,; u_1 u_2 u_3) \\ \mathcal{C}: & \mu(u_1 u_2 r_3, \ldots, u_1 u_2 t_3) > \mu(u_1 u_2 r_0, \ldots, u_1 u_2 t_0\,; u_1 u_2 u_3) \end{array} \tag{82}$$

Prefix various multiples of "r_0 of $\mathtt{a}$ and ... and t_0 of $\mathtt{c}$".

$$\begin{array}{ll} \mathcal{A}: & \mu\big((u_1 u_2 + u_1 u_3) r_0 + u_2 u_3 r_1, \ldots, (u_1 u_2 + u_1 u_3) t_0 + u_2 u_3 t_1\big) < Q \\ \mathcal{B}: & \mu\big((u_1 u_2 + u_2 u_3) r_0 + u_1 u_3 r_2, \ldots, (u_1 u_2 + u_2 u_3) t_0 + u_1 u_3 t_2\big) = Q \\ \mathcal{C}: & \mu\big((u_1 u_3 + u_2 u_3) r_0 + u_1 u_2 r_3, \ldots, (u_1 u_3 + u_2 u_3) t_0 + u_1 u_2 t_3\big) > Q \end{array} \tag{83}$$

where

$$Q = \mu\big((u_1 u_2 + u_1 u_3 + u_2 u_3) r_0, \ldots, (u_1 u_2 + u_1 u_3 + u_2 u_3) t_0\,; u_1 u_2 u_3\big) \tag{84}$$

Evaluate the left-hand sides and eliminate the common right-hand sides Q.

$$\begin{aligned}&\big((u_1u_2+u_1u_3)r_0+u_2u_3r_1\big)a+\cdots+\big((u_1u_2+u_1u_3)t_0+u_2u_3t_1\big)c\\&\quad<\big((u_1u_2+u_2u_3)r_0+u_1u_3r_2\big)a+\cdots+\big((u_1u_2+u_2u_3)t_0+u_1u_3t_2\big)c\\&\qquad<\big((u_1u_3+u_2u_3)r_0+u_1u_2r_3\big)a+\cdots+\big((u_1u_3+u_2u_3)t_0+u_1u_2t_3\big)c\end{aligned}\tag{85}$$

Subtract $(u_1u_2+u_1u_3+u_2u_3)(r_0a+\cdots+t_0c)$ and divide by $u_1u_2u_3$.

$$\begin{aligned}&\underbrace{\big((r_1-r_0)a+\cdots+(t_1-t_0)c\big)\,/\,u_1}_{\text{any member of }\mathcal{A}}\\&\qquad<\underbrace{\big((r_2-r_0)a+\cdots+(t_2-t_0)c\big)\,/\,u_2}_{\text{any member of }\mathcal{B}}\\&\qquad\qquad<\underbrace{\big((r_3-r_0)a+\cdots+(t_3-t_0)c\big)\,/\,u_3}_{\text{any member of }\mathcal{C}}\end{aligned}\tag{86}$$

Taking $((r-r_0)a+\cdots+(t-t_0)c)\,/\,u$ as the statistic, all members of $\mathcal{A}$ lie beneath all members of $\mathcal{B}$, which in turn lie beneath all members of $\mathcal{C}$. We can now assign the value of $\mu(r_0,\ldots,t_0\,;u)$ for some target multiple u. The treatment differs somewhat according to whether or not $\mathcal{B}$ is empty.

A.3.3 *Assignment When $\mathcal{B}$ Has Members*

If $\mathcal{B}$ is non-empty, we now show that all its members share a common value. Let two members be $\{r,\ldots,t\,;u\}$ and $\{r',\ldots,t'\,;u'\}$ (the suffix "2" is temporarily redundant), so that, by definition,

$$\begin{aligned}\mu(r,\ldots,t)&=\mu(r_0,\ldots,t_0\,;u)\\\mu(r',\ldots,t')&=\mu(r_0,\ldots,t_0\,;u')\end{aligned}\tag{87}$$

Apply repetitions by u' and u, respectively.

$$\begin{aligned}\mu(u'r,\ldots,u't)&=\mu(u'r_0,\ldots,u't_0\,;uu')\\\mu(ur',\ldots,ut')&=\mu(ur_0,\ldots,ut_0\,;uu')\end{aligned}\tag{88}$$

Prefix multiples u and u' of "r_0 of $\mathtt{a}$ and ... and t_0 of $\mathtt{c}$".

$$\begin{aligned}\mu(ur_0+u'r,\ldots,ut_0+u't)&=\mu(ur_0+u'r_0,\ldots,ut_0+u't_0\,;uu')\\\mu(u'r_0+ur',\ldots,u't_0+ut')&=\mu(ur_0+u'r_0,\ldots,ut_0+u't_0\,;uu')\end{aligned}\tag{89}$$

Evaluate the left-hand sides and eliminate the common right-hand side.

$$(ur_0+u'r)a+\cdots+(ut_0+u't)c=(u'r_0+ur')a+\cdots+(u't_0+ut')c\tag{90}$$

Subtract $(u+u')(r_0a+\cdots+t_0c)$ and divide by uu',

$$\frac{(r-r_0)a+\cdots+(t-t_0)c}{u}=\frac{(r'-r_0)a+\cdots+(t'-t_0)c}{u'}=d\tag{91}$$

in which d denotes this common value now seen to be shared by all members of $\mathcal{B}$. Using the definitions again, evaluating, and using this common value gives

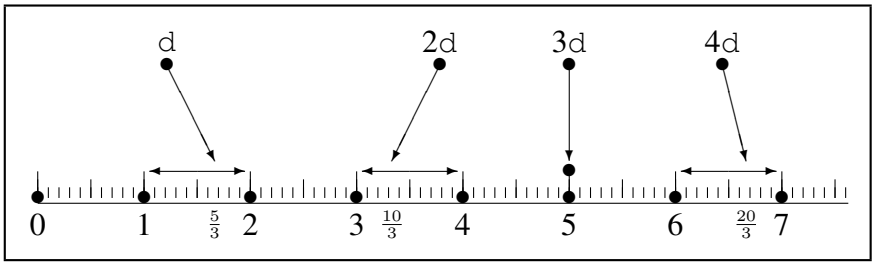
$$\mu(r_0,\ldots,t_0\,;u)=\mu(r,\ldots,t)=ra+\cdots+tc=r_0a+\cdots+t_0c+ud\tag{92}$$

where d is seen to be the value $m(\mathtt{d})$ of a single atom of type $\mathtt{d}$. By Equation 91, this value is rationally related to the previous values $a, \ldots, c$.

Illustration

Suppose for simplicity that only one type of atom has previously been assigned ($k = 1$), according to the integer scale $m(r \text{ of } \mathtt{a}) = ra$ with $a = 1$. Suppose that the new atom $\mathtt{d}$ has value $d = \frac{5}{3}$, rationally related to a. By 3-fold repetition, this means that $m(3 \text{ of } \mathtt{d})$ lies exactly at 5, and is a member of set $\mathcal{B}$. Again by 3-fold repetition, $m(1 \text{ of } \mathtt{d})$ cannot lie at or below 1 because that would wrongly imply $m(3 \text{ of } \mathtt{d}) \leq 3$. Similarly, it cannot lie at or above 2 because that would imply $m(3 \text{ of } \mathtt{d}) \geq 6$. So $m(1 \text{ of } \mathtt{d})$ necessarily lies between 1 (which lies in set $\mathcal{A}$) and 2 (which lies in set $\mathcal{C}$) and can without loss of generality be assigned $\frac{5}{3}$. Similarly, $m(2 \text{ of } \mathtt{d})$ necessarily lies between 3 and 4 and can without loss of generality be assigned $\frac{10}{3}$, and so on (Figure 8).

Figure 8. Multiples of a new type of atom can be assigned linear values.



These assignments obey axioms 1 and 2, and we now have $\mathtt{a}$ and $\mathtt{d}$ on the *same* linear scale.

A.3.4 *Assignment When $\mathcal{B}$ Has no Members*

When $\mathcal{B}$ is empty, the strict inequalities Equation 86 separating $\mathcal{A}$ and $\mathcal{C}$ imply that partitioning between them can be accomplished by some real δ.

$$\underbrace{\frac{(r_1 - r_0)a + \cdots + (t_1 - t_0)c}{u_1}}_{\text{any member of } \mathcal{A}} < \delta < \underbrace{\frac{(r_3 - r_0)a + \cdots + (t_3 - t_0)c}{u_3}}_{\text{any member of } \mathcal{C}} \tag{93}$$

For the target multiplicity u, the definitions Equation 80 showed $\mu(r_0, \ldots, t_0\,; u)$ to be bounded below by those members of $\mathcal{A}$ having $u_1 = u$, and bounded above by those members of $\mathcal{C}$ having $u_3 = u$. These constraints relevant to the target u are

$$\underbrace{r_1 a + \cdots + t_1 c}_{\text{subset } u_1 = u \text{ of } \mathcal{A}} < \mu(r_0, \ldots, t_0\,; u) < \underbrace{r_3 a + \cdots + t_3 c}_{\text{subset } u_3 = u \text{ of } \mathcal{C}} \tag{94}$$

which is equivalent to

$$\underbrace{\frac{(r_1-r_0)a+\cdots+(t_1-t_0)c}{u}}_{\text{subset } u_1=u \text{ of } \mathcal{A}} < \frac{\mu(r_0,\ldots,t_0\,;u)-(r_0a+\cdots+t_0c)}{u} < \underbrace{\frac{(r_3-r_0)a+\cdots+(t_3-t_0)c}{u}}_{\text{subset } u_3=u \text{ of } \mathcal{C}} \tag{95}$$

Because this refers to subsets involving a single u rather than the entire sets involving all u, it is a weaker constraint than the preceding global constraint Equation 93 was on δ. In other words, its central quantity $\big(\mu(r_0,\ldots,t_0\,;u)-(r_0a+\cdots+t_0c)\big)\,/\,u$ is allowed to lie anywhere within an interval that contains the narrower interval containing δ. Accordingly, it is legitimate to assign

$$\frac{\mu(r_0,\ldots,t_0\,;u)-(r_0a+\cdots+t_0c)}{u}=\delta \tag{96}$$

which automatically satisfies all the relevant constraints Equation 95. So the simple assignment

$$\mu(r_0,\ldots,t_0\,;u)=r_0a+\cdots+t_0c+u\delta \tag{97}$$

automatically falls in the correct interval. The only freedom is regrade to some alternative value within the relevant interval.

Illustration

Suppose that three types of atom have previously been assigned ($k=3$), according to

$$m(r \text{ of } \mathtt{a}) \oplus m(s \text{ of } \mathtt{b}) \oplus m(t \text{ of } \mathtt{c}) = ra+sb+tc \tag{98}$$

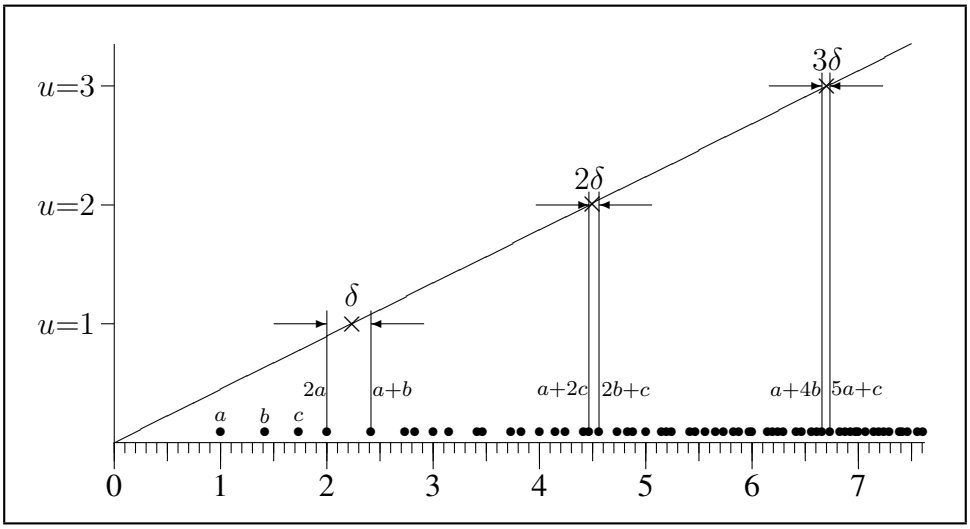
with $a=1$, $b=\sqrt{2}$, $c=\sqrt{3}$. Now introduce a fourth type $\mathtt{d}$. Omitting r_0, s_0, t_0 for simplicity, we might find that multiples u of $\mathtt{d}$ fall into successive intervals as follows.

$$\begin{array}{rcr}
2.0000=2a & <m(1 \text{ of } \mathtt{d})< & a+b= \;\; 2.4142\\
4.4641=a+2c & <m(2 \text{ of } \mathtt{d})< & 2b+c= \;\; 4.5605\\
6.6569=a+4b & <m(3 \text{ of } \mathtt{d})< & 5a+c= \;\; 6.7321\\
 & \cdots & \\
22.3424=14a+b+4c & <m(10 \text{ of } \mathtt{d})< & 9a+7b+2c=22.3636\\
 & \cdots &
\end{array} \tag{99}$$

These are the constraints Equation 94 relevant to each individual $u=1,2,3,\ldots,10,\ldots$. It is guaranteed that there exists some δ such that the relevant interval for each target multiplicity u covers $u\delta$, as illustrated by the diagonal line of slope $1/\delta$ in the diagram. Any breakout from these intervals would have contradicted axiom 2 thereby showing that δ had been incorrectly assigned (Figure 9).

According to Equation 93 with $r_0=s_0=t_0=0$, the value of $\delta=m(u \text{ of } \mathtt{d})\,/\,u$ is constrained by all the members of $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$.

Figure 9. Multiples of a new atom can always be assigned linear values $\delta, 2\delta, 3\delta, \ldots$. An individual multiple can be assigned anywhere within the corresponding interval, but the linear assignment can always be chosen.



By the time these sets have expanded to cover up to 10 copies of $\mathtt{d}$, the surviving interval is

$$2.2360 = \underbrace{(8a+7c)/9}_{\text{from } u_1=9} < \delta < \underbrace{(7a+5c)/7}_{\text{from } u_3=7} = 2.2372 \tag{100}$$

and by the time 1000 copies are allowed, the union of all the constraints fixes δ to 10 decimal places.

$$2.236067977497 = \underbrace{\frac{1345a+56b+359c}{915}}_{\text{from } u_1=915} < \delta < \underbrace{\frac{80a+545b+286c}{602}}_{\text{from } u_3=602} = 2.236067977505 \tag{101}$$

(The example happened to have $\delta = \sqrt{5}$.)

Accuracy

The gap between $\mathcal{A}$ and $\mathcal{C}$ might allow δ to be uncertain. We assume that δ is bounded below, otherwise the appended atoms of type $\mathtt{d}$ never have measurable effect. This implies the existence of u such that $u\delta > na$ for any multiple n, no matter how large. We also assume that δ is bounded above, otherwise even a single $\mathtt{d}$ atom always overwhelms everything else. This implies the existence of a greatest $r \geq n$ such that $ra < u\delta$ for that u. Taking other types of atom to be absent for simplicity, we have

$$(r, 0, \ldots, 0\,; u) \in \mathcal{A} \quad \text{and} \quad (r+1, 0, \ldots, 0\,; u) \in \mathcal{C} \tag{102}$$

where r can be indefinitely large. The corresponding inequalities $ra/u < \delta < (r+1)a/u$ from (93) fix δ to accuracy 1 part in r (1 in n or better).

This proves that δ can be found to arbitrarily high accuracy by allowing sufficiently high multiples. Denote the limiting value of δ by d. This value $m(\mathtt{d}) = d$ of a single atom of type $\mathtt{d}$ is now fixed to unlimited accuracy, but has no rational relationship to the previous values $a, \ldots, c$.

A.3.5 *End of Inductive Proof*

Whether or not $\mathcal{B}$ had members, the assignment

$$\mu(r_0, \ldots, t_0\,; u) = r_0 a + \cdots + t_0 c + ud \tag{103}$$

obeys all the defining inequalities Equation 80. This updates the original assignment Equation 71 from k atom types to $k+1$, so by induction from $k = 1$ it holds for any k.

$$\underbrace{m(r \text{ of } \mathtt{a} \text{ and} \ldots \text{ and } t \text{ of } \mathtt{c} \text{ and} \ldots \text{ and } v \text{ of } \mathtt{e})}_{\text{any number of types in any order}} = \underbrace{ra + \cdots + tc + \cdots + ve}_{\text{corresponding terms}} \tag{104}$$

Atom types in the above expression are often different, but do not need to be, and the formula represents the quantification of a general sequence. Embedded in it, and equivalent to it, is the sum rule $x \oplus y = x + y$ for the values $m(\mathtt{x}) = x$ and $m(\mathtt{y}) = y$ of arbitrary sequences. Any order-preserving regrade Θ is also permitted, but no order-breaking transform is permitted.

This completes the inductive proof for atoms of positive style. The proof holds equally well for atoms of negative style, for which the values are negative. Meanwhile, Equation 68 shows that atoms of null style have zero value. So, even if the atoms may have arbitrary style, Equation 104 offers the only consistent combination rule. The result thus holds for atom values of arbitrary sign and arbitrary magnitude, though the nature of the constructive proof requires atom *multiplicities* to be non-negative. □

A.4. Axioms are Minimal

Theorem:

Axioms 1a, 1b, 2 are individually required.

Proof:

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

Without axiom 1a (postfix ordering), the definition

$$a \bigcirc b = \lfloor a \rfloor + b \tag{105}$$

where $\lfloor a \rfloor$ is the integer at or immediately below a, satisfies axioms 1b and 2 but cannot be equivalent to addition because it is not commutative; $a \bigcirc b \neq b \bigcirc a$. So axiom 1a is required.

Without axiom 1b (prefix ordering), the definition

$$a \bigcirc b = a + \lfloor b \rfloor \tag{106}$$

satisfies axioms 1a and 2, but cannot be equivalent to addition because it is not commutative. So axiom 1b is required.

Without axiom 2 (associativity), the definition

$$x \bigcirc y = x^2 + y^2 \tag{107}$$

satisfies axioms 1a and 1b (ordering), and also happens to be continuous and commutative ($x \bigcirc y = y \bigcirc x$). Yet it cannot be equivalent to addition because $\Theta(x \bigcirc y) = \Theta(x) + \Theta(y)$ has no solution that would enable a regrade Θ. That can be shown by appropriately differencing $\delta_x \delta_y$ to reach $\Theta(z+\epsilon) - 2\Theta(z) + \Theta(z-\epsilon) = 0$ whose solution $\Theta(z) = Az + B$ fails to satisfy the supposedly defining Equation 107. Hence ordering is insufficient even when accompanied by continuity and commutativity. Axiom 2 (associativity) is definitely required. □

B. Appendix B: Product Theorem

Theorem:

The solution of the functional **product Equation**

$$\Psi(\tau + \xi) + \Psi(\tau + \eta) = \Psi\big(\tau + \zeta(\xi, \eta)\big) \tag{108}$$

in which τ, ξ and η are independent real variables and Ψ is positive is

$$\Psi(x) = Ce^{Ax} \tag{109}$$

where A and C are constants (C necessarily being positive).

B.1. **Proof:**

The quoted solution is easily seen to satisfy the product equation, which demonstrates *existence*. The remaining question is whether the solution is *unique*.

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

$$2\Psi(\tau + \zeta - a) = \Psi(\tau + \zeta) \tag{110}$$

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

$$\Psi(\theta + na) = 2^n \Psi(\theta) \tag{111}$$

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

$$\Psi(\tau + \zeta - b) + \Psi(\tau + \zeta - 2b) = \Psi(\tau + \zeta) \tag{112}$$

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

$$\Psi(\theta+mb)=\left(\frac{2\,\Psi(\theta)}{5+\sqrt{5}}+\frac{\Psi(\theta+b)}{\sqrt{5}}\right)\left(\frac{1+\sqrt{5}}{2}\right)^m+\left(\frac{2\,\Psi(\theta)}{5-\sqrt{5}}-\frac{\Psi(\theta+b)}{\sqrt{5}}\right)\left(\frac{-2}{1+\sqrt{5}}\right)^m \tag{113}$$

for any integer m, θ being arbitrary.

This combines with the 2-term formula to make

$$\begin{aligned}\Psi(\theta+mb-na)=&\left(\frac{2\,\Psi(\theta)}{5+\sqrt{5}}+\frac{\Psi(\theta+b)}{\sqrt{5}}\right)e^{m\log\left(\frac{1+\sqrt{5}}{2}\right)-n\log 2}+\\&(-1)^m\left(\frac{2\,\Psi(\theta)}{5-\sqrt{5}}-\frac{\Psi(\theta+b)}{\sqrt{5}}\right)e^{-m\log\left(\frac{1+\sqrt{5}}{2}\right)-n\log 2}\end{aligned} \tag{114}$$

For any integer n, there is an even integer m for which $0\le mb-na<2b$ 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

$$\Psi(\theta+mb-na)=\Psi(\theta)\,e^{m\log\left(\frac{1+\sqrt{5}}{2}\right)-n\log 2} \tag{115}$$

(first term only) or

$$\Psi(\theta+mb-na)=(-1)^m\Psi(\theta)\,e^{-m\log\left(\frac{1+\sqrt{5}}{2}\right)-n\log 2} \tag{116}$$

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

$$\frac{b}{a}=\frac{\log\left(\frac{1+\sqrt{5}}{2}\right)}{\log 2} \tag{117}$$

and in the second case, bounded Ψ requires

$$\frac{b}{a}=-\frac{\log\left(\frac{1+\sqrt{5}}{2}\right)}{\log 2} \tag{118}$$

Either way,

$$\Psi(\theta+mb-na)=\Psi(\theta)\,e^{A(mb-na)} \tag{119}$$

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

$$\Psi(x)=e^{A(mb-na)}\Psi(\epsilon)=e^{A(x-\epsilon)}\Psi(\epsilon)\approx e^{Ax}\Psi(\epsilon) \tag{120}$$

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

$$\text{(solution)} \qquad \Psi(x) = Ce^{Ax} \tag{121}$$

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

This obeys the original product equation without further restriction and 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 $b = A^{-1} \log(\frac{1+\sqrt{5}}{2})$ were appropriate constants. □

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

$$\text{(inverse)} \qquad \Theta(u) = \frac{1}{A} \log \frac{u}{C} \tag{122}$$

in which u and hence C are both positive.

C. Appendix C: Variational Theorem

Theorem:

The solution of the functional **variational equation**

$$H'(m_x m_y) = \lambda(m_x) + \mu(m_y) \tag{123}$$

with positive m_x and m_y is

$$H(m) = A + Bm + C(m \log m - m) \tag{124}$$

where A, B, C are constants.

C.1. **Proof:**

The quoted solution is easily seen to satisfy the variational equation, with corollaries that the functions λ and μ are logarithmic, which demonstrates *existence*. The remaining question is whether the solution is *unique*.

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

$$h(u + v) = \lambda^*(u) + \mu^*(v) \tag{125}$$

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

$$h(u + v) = h(u) + h(v) - B \tag{126}$$

This is Cauchy's functional equation ([14])

$$f(u + v) = f(u) + f(v) \tag{127}$$

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

$$f(t) = ct \tag{128}$$

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

$$\Phi(u, v) = \iint dx\, dy\, \frac{\mathsf{1}(x^2 + xy + y^2 < \frac{3}{4}\epsilon^2)}{\sqrt{3}\pi\epsilon^2/2}\, \phi(u - x, v - y) \tag{129}$$

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

$$F(u + v) = F(u) + F(v) \tag{130}$$

as before, with the new function

$$F(t) = \int_{-\epsilon}^{\epsilon} dx\, \frac{2\sqrt{\epsilon^2 - x^2}}{\pi\epsilon^2}\, f(t - x) \tag{131}$$

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

$$F(t) = Ct \tag{132}$$

can only have one value for the constant C.

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

$$\begin{aligned}
H(m) &= Bm + \int^{m} f(\log m')dm' && \text{(integrate)}\\
&= Bm + \int^{\log m} f(t)\, e^t dt && \text{(change variable)}\\
&= Bm + \int_{-\epsilon}^{\epsilon} dx\, \frac{2\sqrt{\epsilon^2 - x^2}}{\pi\epsilon^2} \int^{\log m} f(t)\, e^t dt && \text{(insert blurring)}\\
&= Bm + \int_{-\epsilon}^{\epsilon} dx\, \frac{2\sqrt{\epsilon^2 - x^2}}{\pi\epsilon^2} \int^{x+\log m} f(t - x)\, e^{t-x} dt && \text{(offset dummy } t\text{)}\\
&\approx Bm + \int_{-\epsilon}^{\epsilon} dx\, \frac{2\sqrt{\epsilon^2 - x^2}}{\pi\epsilon^2} \int^{\log m} f(t - x)\, e^t dt && (|x| \le \epsilon \text{ small})\ !\\
&= Bm + \int^{\log m} F(t)\, e^t dt && \text{(definition of } F\text{)}\\
&= Bm + C \int^{\log m} te^t dt && \text{(substitute)}
\end{aligned} \tag{133}$$

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

$$H(m) = A + Bm + C(m \log m - m)\,. \tag{134}$$

This obeys the original variational equation with corollaries $\lambda(x) = B_1 + C\log(x)$ and $\mu(x) = B_2 + C\log(x)$ where $B_1 + B_2 = B$, and is the general solution. □

© 2012 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/3.0/).