Algebraic Foundations of Inference: From Order and Symmetry to Probability and Entropy

Codex 5.2, Claude 4.5, Zar Goertzel*

January 26, 2026

Abstract

This note gives a mathematical exposition of an algebraic route from order and symmetry principles to the standard probability calculus and to information measures such as Kullback–Leibler divergence and Shannon entropy.[1]

The emphasis is on the *minimal mathematical hypotheses* needed for each step, and on how the resulting axiom systems compare with classical foundations (Kolmogorov, Cox) and with the Shore–Johnson consistency axioms for maximum entropy.

Our formal proofs live in Lean 4; Lean details are relegated to footnotes so the main text reads as ordinary mathematics.

Contents

*This document was drafted collaboratively by humans and AI systems (GPT/Codex, Claude). While formal claims are machine-checked in Lean 4, prose descriptions may contain human or AI hallucinations. Caveat lector.

[1]This development is based on a formalization of Knuth & Skilling [1], which we have strengthened and present here in a self-contained way.

1 Introduction: why an algebraic foundation?

Probability theory is often presented in two complementary foundational styles:

- **Measure-theoretic** foundations (Kolmogorov): start with a σ-algebra of events and a probability measure P satisfying axioms of countable additivity.
- **Plausibility-theoretic** foundations (Cox and successors): start with a notion of rational plausibility for propositions and derive the probability calculus from consistency desiderata.

An algebraic viewpoint is that inference is about *combining* information, and combination operations have symmetries (associativity, distributivity, etc.). These symmetries constrain the possible numerical representations so strongly that familiar calculi emerge.[2]

In the spirit of reverse mathematics, we isolate necessary hypotheses for each representation theorem (additive, multiplicative, conditional, variational) and explore which combinations of axioms yield which conclusions. We also make "regularity gates" explicit wherever functional equations admit pathological solutions.

Roadmap (what implies what)

The algebraic program can be read as a sequence of increasingly structured "symmetry laws" and their consequences:

- **Disjoint join.** Associativity + strict order-compatibility + an Archimedean/density axiom $\Rightarrow$ regrade $\oplus$ to real addition (section 4).
- **Independent product.** Distributivity of $\otimes$ over $+$ + associativity of $\otimes \Rightarrow \otimes$ is multiplication up to a constant scale (section 5).
- **Conditioning.** A chaining axiom for $p(a \mid b)$ on a lattice of events $\Rightarrow$ Bayes/product rules (section 6).

[2] See [1] for an influential instance of this viewpoint.

- **Variation.** A separated variational functional equation for a potential H + an explicit regularity gate $\Rightarrow$ the logarithmic normal form and thus KL divergence and Shannon entropy (section 7).
- **Extension to σ-additivity.** Completeness axioms on scale and events $\Rightarrow$ σ-additivity as a theorem, with a bridge to measure theory (section 8).

Axiom summary (at a glance)

Hypotheses	**Conclusion**
Ordered scale $(S, \oplus)$ + no anomalous pairs	$\Theta(x \oplus y) = \Theta(x) + \Theta(y)$
Additive coordinate + distributive/associative $\otimes$	$x \otimes y = xy/C$ (scaled multiplication)
Distributive event lattice + chaining law for $p(a \mid b)$	product rule and Bayes' rule
Variational equation + Gate R (regularity)	$H'(m) = B + C \log m$ and KL/entropy forms
Finite additivity + scale/event completeness	σ-additivity (measure theory bridge)

The two gates: globality and regularity

Two recurring *hypothesis gates* appear throughout the algebraic program. Understanding them is essential.

Gate G (Globality / Richness). When can a local optimality condition (at a stationary point) be upgraded to a *global* functional equation? This requires a richness/universality premise: that the class of admissible problems is broad enough to test the functional constraint at essentially arbitrary inputs.

Where it applies: The *variational theorem* (section 7). The entropy functional equation $H'(m_x m_y) = \lambda(m_x) + \mu(m_y)$ arises from Lagrange multiplier conditions at stationary points; Gate G justifies extending this to all positive (m_x, m_y).

Gate R (Rigidity / Regularity). When does an algebraic functional equation have a unique "nice" solution? Cauchy-type equations $f(x + y) = f(x) + f(y)$ and their multiplicative variants admit pathological solutions (Hamel bases) unless a regularity hypothesis is imposed: measurability, monotonicity, or continuity at a point suffice; see standard references such as [8, 9].

Where it applies:

- *Product representation* (section 5): The step "additive on $(0, \infty)$ implies linear" is a Cauchy-type equation. Here, **positivity** of $\otimes$ supplies the needed monotonicity, so Gate R is satisfied implicitly.
- *Variational theorem* (section 7): The equation $H'(m_x m_y) = \lambda(m_x) + \mu(m_y)$ reduces to Cauchy's equation after a log-coordinate change. Here, Gate R must be imposed explicitly (e.g., H' is measurable or monotone).

Throughout this paper, we keep both gates explicit rather than sweeping them under implicit "reasonableness" assumptions.

2 Events, order, and the idea of regraduation

The common core behind plausibility-theoretic and algebraic foundations is that propositions (or "events") carry an order: one event may entail another. Let E be a partially ordered set of events with bottom $\bot$ (false) and top $\top$ (true).

Definition 2.1. A *valuation* is a map $v : E \to S$ into a totally ordered *scale* S such that $a \leq b \Rightarrow v(a) \leq v(b)$.

Different choices of scale S and of numerical coordinate on S are often physically irrelevant. A change of scale is a *regraduation*.[3] Mathematically, it is an order isomorphism $\Theta : S \to S'$.

Remark 2.2 (Order-preserving regraduations are rigid once additivity is fixed). A recurring subtlety in algebraic foundations is that *functional equations have pathological solutions* unless one imposes a "regularity gate". For example, the additive Cauchy equation admits wildly discontinuous solutions (via Hamel bases). However, once one requires *order preservation* (fidelity/monotonicity), these pathologies disappear: any monotone additive map on $\mathbb{R}$ is necessarily linear, hence continuous (see e.g. [8, 9]). In other words: once the sum rule has been regraded to real addition, the only coordinate changes that preserve both the sum rule and order are affine transformations.

Remark 2.3 (Why order is the right primitive). Order captures the idea of *fidelity*: if a entails b, then a should not be assigned a larger plausibility than b. Many apparently "numerical" axioms in probability can be viewed as order axioms plus structural symmetries of event-combination operations.

3 Event structures: Boolean and beyond

For finite reasoning it is convenient to work with a lattice of events. Write $a \vee b$ for join (logical OR) and $a \wedge b$ for meet (logical AND).

Definition 3.1. A *distributive lattice* is a set E equipped with $\vee, \wedge$ such that both operations are associative, commutative, idempotent, and satisfy absorption, and such that $\wedge$ distributes over $\vee$ (equivalently $\vee$ distributes over $\wedge$).

Definition 3.2. Two events $a, b \in E$ are *disjoint* if $a \wedge b = \bot$. In a Boolean algebra, this is equivalent to $a \leq \neg b$, but in general we do not assume complements exist.

A key algebraic point is that complements are not primitive: the sum rule and product rule can be developed in non-Boolean distributive lattices.[4]

Example 3.3 (Boolean algebra of subsets). For a finite set Ω, the powerset $\mathcal{P}(\Omega)$ is a Boolean algebra with $\vee = \cup$, $\wedge = \cap$, $\bot = \emptyset$, $\top = \Omega$.

Example 3.4 (Three-element chain (minimal non-Boolean distributive lattice)). Let $E = \{\bot < a < \top\}$ with $\vee = \max$, $\wedge = \min$. This lattice is distributive but not Boolean: a has no complement in E.

Example 3.5 (Open sets). For a topological space X, the collection of open sets is a distributive lattice under $\cup, \cap$, typically not Boolean because complements of open sets need not be open.

[3] The term is used prominently in [1].

[4] This is emphasized in [1], but the underlying observation is independent of any particular presentation.

PART I: ABSTRACT REPRESENTATION THEOREMS

The next three sections develop the abstract mathematics of ordered semigroups, tensors, and functional equations—independent of any particular event interpretation.

4 From associative combination to addition (additive representation)

4.1 The algebraic scale for disjoint combination

Introduce a binary operation $\oplus$ on scale values to represent the valuation of a disjoint join: if a and b are disjoint events, then $v(a \vee b)$ is determined by $v(a)$ and $v(b)$, so

$$v(a \vee b) = v(a) \oplus v(b).$$

Abstractly, the scale $(S, \oplus)$ is required to satisfy:

- **Associativity:** $(x \oplus y) \oplus z = x \oplus (y \oplus z)$.
- **Strict monotonicity in each argument:** $x < y \Rightarrow x \oplus z < y \oplus z$ and $z \oplus x < z \oplus y$.

The algebraic core is identity-free. The two axioms above (associativity and strict monotonicity) constitute the minimal algebraic structure for an ordered semigroup. Classical representation theory (Alimov [10], Hölder [12]) works at this level of generality: no identity element is required.

Probability semantics: identity as minimum (optional add-on). For probability-style applications, one often adds:

- **Identity:** an element 0 such that $x \oplus 0 = x = 0 \oplus x$.
- **Minimum:** $0 \leq x$ for all x.

The identity provides a canonical normalization $\Theta(0) = 0$ for the representation. The minimum condition rules out "negative" anomalous behavior automatically.

Definition 4.1 (Iteration / "n of x")**.** Fix a scale $(S, \oplus, 0)$. Define the iterates

$$x^{(0)} := 0, \qquad x^{(n+1)} := x \oplus x^{(n)}.$$

When $\oplus$ is (regraded to) real addition, $x^{(n)}$ is just $n \cdot x$.

4.2 The density axiom: no anomalous pairs (Alimov–Hölder)

Associativity and monotonicity are not enough to force an additive real representation. One also needs an *Archimedean/density* condition ruling out infinitesimal gaps.

The classical formulation, due to Alimov [10] building on Hölder [12], is the *no anomalous pairs* condition.

Definition 4.2 (Anomalous pair)**.** Say that $a, b > 0$ form an *anomalous pair* if for every $n \geq 1$,

$$a^{(n)} < b^{(n)} < a^{(n+1)}.$$

Intuitively, b is "infinitesimally larger" than a: no finite magnification makes the gap exceed a.

Axiom 4.3 (No Anomalous Pairs (NAP))**.** No pair of positive elements forms an anomalous pair.

Theorem 4.4 (Alimov–Hölder representation)**.** *In a linearly ordered cancellative semigroup, NAP is equivalent to embeddability into* $(\mathbb{R}, +)$*.*

This provides a clean conceptual bridge: "one-dimensionality" is exactly the exclusion of infinitesimals.[5] Our project uses Eric Luap's Lean formalization of Hölder's theorem for this route.

Remark 4.5 (What NAP excludes—necessity witness). NAP excludes lexicographic "multi-scale" structures with infinitesimals. **Example (lexicographic structure):** $(\mathbb{N} \times \mathbb{N}, +)$ ordered lexicographically admits anomalous pairs: if $a = (1, 0)$ and $b = (1, 1)$, then $a^{(n)} = (n, 0)$ and $b^{(n)} = (n, n)$, so $a^{(n)} < b^{(n)} < a^{(n+1)}$ for all n.

4.3 Equivalent formulation: separation (Claude Code, Ben Goertzel)

An equivalent formulation of the density axiom, developed collaboratively by Claude Code and Ben Goertzel during this project, is *separation*—a direct rational approximation principle.

Axiom 4.6 (Separation)**.** For any $0 < a$ and any $0 < x < y$, there exist natural numbers $m > 0, n$ such that

$$x^{(m)} \;<\; a^{(n)} \;\leq\; y^{(m)},$$

where $x^{(m)}$ denotes m-fold iteration of $\oplus$.

Theorem 4.7 (Separation $\Leftrightarrow$ NAP (bounded-below setting))**.** *In a linearly ordered monoid where the identity* 0 *is the minimum:*

1. *Separation $\Rightarrow$ NAP.*
2. *NAP $\Rightarrow$ (via Hölder representation) $\Rightarrow$ Separation (by rational density in $\mathbb{R}$).*

Theorem 4.8 (Bilateral Separation $\Leftrightarrow$ NAP)**.** *In a linearly ordered monoid with strict monotonicity (without assuming identity is minimum), define* bilateral separation*:*

- ***(Positive)** For $0 < a$ and $0 < x < y$: $\exists\, m > 0$, $n \geq 0$ with $x^{(m)} < a^{(n)} \leq y^{(m)}$.*
- ***(Negative)** For $a < 0$ and $x < y < 0$: $\exists\, m > 0$, $n \geq 0$ with $x^{(m)} < a^{(n)} \leq y^{(m)}$.*

Then: ***NAP $\Leftrightarrow$ Bilateral Separation.***

The key constraint is that a, x, y must lie on the same side *of identity; cross-sign sandwiching is impossible since rationals cannot approximate sign-mismatched ratios.*

Theorem 4.9 (Separation forces commutativity)**.** *In a linearly ordered scale with strict monotonicity, separation implies:*

1. ***Archimedean growth:** for any $0 < a$ and any x, some iterate $a^{(n)}$ exceeds x;*
2. ***Commutativity:** $x \oplus y = y \oplus x$ (so commutativity need not be assumed).*

Remark 4.10. In probability applications, identity represents impossibility ($\perp$) and is the minimum of the order, so the negative branch of bilateral separation is vacuously satisfied.

[5] A key connection: in an ordered semigroup with strict monotonicity, non-commutativity forces the existence of anomalous pairs. Contrapositive: NAP implies commutativity, which we prove as theorem 4.9.

4.4 Representation theorem: the additive coordinate

The core conclusion is:

Theorem 4.11 (Additive representation)**.** ***Assumptions:***

- $(S, \oplus)$ *is a linearly ordered semigroup with strict monotonicity in each argument.*
- *No anomalous pairs (and, when a minimum identity exists, equivalently: separation).*
- *Optional: identity element* 0 *with* $0 \leq x$ *for all* x *(for probability semantics).*

Conclusion: *There exists an order embedding* $\Theta : S \to \mathbb{R}$ *such that*

$$\Theta(x \oplus y) = \Theta(x) + \Theta(y).$$

The representation is unique up to positive affine transformation. If identity exists, one can normalize $\Theta(0) = 0$*.*

We have three proof routes with different hypothesis profiles:

1. **Hölder embedding (main route)**: classical ordered-semigroup representation from NAP. This is the conceptually cleanest path, working on identity-free semigroups.
2. **Dedekind cuts**: classical density construction from separation.
3. **Grid induction**: explicit finite construction (more constructive, more complex).

5 From distributive tensors to multiplication (product representation)

Introduce a second operation $\otimes$ to represent the valuation of independent products (direct products of lattices / independent systems). After regraduating $\oplus$ to real addition, the key axiom is distributivity:

$$(x + y) \otimes t = (x \otimes t) + (y \otimes t).$$

Remark 5.1 (Why the multiplicative case needs a different interface). For multiplication on $(0, \infty)$, the identity 1 is *not* a minimum. So an attempt to reuse the additive semigroup interface for $\otimes$ would be structurally mismatched. Instead one derives that $\otimes$ is multiplication-like from distributivity and associativity.

Theorem 5.2 (Product representation)**.** ***Assumptions:***

- $(0, \infty)$ *with real addition (from the additive representation theorem).*
- *A binary operation* $\otimes : (0, \infty)^2 \to (0, \infty)$ *that is:*
  - *Distributive over* $+$*:* $(x + y) \otimes t = (x \otimes t) + (y \otimes t)$.
  - *Associative:* $(x \otimes y) \otimes z = x \otimes (y \otimes z)$.
  - *Positive:* $x, y > 0 \Rightarrow x \otimes y > 0$.

Conclusion: *There exists a constant* $C > 0$ *such that* $x \otimes y = xy/C$.

We have two proof routes with different flavors.

5.1 Direct proof (algebraic)

Fix t and consider $f_t(x) := x \otimes t$ on $(0, \infty)$. Distributivity says f_t is additive on positive reals. Positivity of $\otimes$ rules out Hamel-basis pathologies (Gate R) and forces $f_t(x) = k(t)\, x$. Associativity then forces the scale factors to be coherent: $k(t) = c\, t$ for a constant $c > 0$, yielding $x \otimes y = xy/C$.

5.2 Fibonacci proof (K&S original)

K&S's Appendix B solves a functional equation $\Psi(\xi + \tau) + \Psi(\eta + \tau) = \Psi(\zeta(\xi, \eta) + \tau)$ by deriving recurrence relations. A 2-term recurrence gives $\Psi(\theta + na) = 2^n \Psi(\theta)$; a 3-term recurrence yields the golden ratio φ. The irrationality of φ implies that offsets $m \cdot \log \varphi - n \cdot \log 2$ are dense in $\mathbb{R}$, forcing $\Psi(x) = Ce^{Ax}$—an exponential, hence multiplication after a log-coordinate change.

Remark 5.3 (Assumptions in the direct proof). The Cauchy-type step "additive on $(0, \infty)$ implies linear" needs an anti-pathology hypothesis. In the direct proof, *positivity* of $\otimes$ supplies the needed monotonicity (Gate R). The final step (from $f_t(x) = k(t)x$ to $k(t) = ct$) requires only injectivity of k. The Fibonacci proof uses the same positivity assumption but via a different route (irrationality of the golden ratio forces the exponential form).

Part II: Interpretation and Applications

Having established the abstract representation theorems, we now interpret them in the context of event lattices, conditional plausibility, and information measures.

6 Conditional plausibility on lattices

Complements are not primitive: inference can be carried out on distributive lattices that are not Boolean. One can axiomatize a conditional plausibility $p(a \mid b)$ on comparable event pairs with a chaining/product axiom and derive Bayes-style identities.[6]

The statements below are marked "schematic" because the full type-theoretic details (handling of partiality, domain constraints, etc.) are in the Lean formalization; here we present the conceptual content.

Definition 6.1 (Conditional plausibility (schematic))**.** Given events $a \leq b$, write $p(a \mid b)$ for the plausibility of a in the context b. One requires a context law $p(a \mid b) = p(a \wedge b \mid b)$ and monotonicity in a.

Axiom 6.2 (Chaining / product axiom (schematic))**.** For $a \leq b \leq c$ with $a \neq \bot$, there is a binary operation $\star$ on plausibilities such that

$$p(a \mid c) = p(a \mid b) \star p(b \mid c).$$

Theorem 6.3 (Sum, product, Bayes (schematic))**.** *Under the additive and multiplicative representations above, the chaining axiom yields the familiar probability calculus:*

$$P(A \wedge B) = P(A \mid B)\, P(B), \qquad P(A \mid B)\, P(B) = P(B \mid A)\, P(A).$$

The key insight is that the chaining axiom does not assume $\star$ is multiplication—only that *some* operation chains conditionals. The algebraic constraints (associativity of $\star$, compatibility with the product rule for independent events) then *force* $\star$ to be multiplication. It is not a choice; it is the unique coherent possibility.

7 Variational potentials, divergence, and entropy

The pattern continues: just as associativity + order forced $\oplus$ to be addition, and distributivity + associativity forced $\otimes$ to be multiplication, a "product-to-sum" separation condition on a variational potential forces the *logarithmic* form—yielding KL divergence and Shannon entropy as the unique information measures consistent with the algebraic structure.

Concretely, introduce a potential H whose derivative satisfies a separated functional equation

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

After the log-coordinate change $u = \log m$, this becomes Cauchy's additive equation.

[6] This "chaining" symmetry is central in [1], but it can be stated without reference to any particular presentation.

7.1 Two independent regularity issues

1. **(Globality)** Turning a local Lagrange-multiplier separation (at a stationary point) into a global functional equation requires an explicit universality/richness premise.

2. **(Anti-pathology)** Solving the Cauchy equation uniquely requires a regularity gate (measurable / monotone / continuous); without it, Hamel-basis solutions exist.

Our development keeps both premises explicit and provides counterexamples showing why some anti-pathology hypothesis is logically necessary.

Definition 7.1 (Universality and richness (schematic))**.** "Universality across applications" means a single potential H is intended to apply across a wide class of inference problems, not tuned to a specific domain. "Richness" means that admissible problems range widely enough that the functional constraints implied by local optimality can be tested at essentially arbitrary positive pairs (m_x, m_y). Together, these premises justify upgrading a local separation condition to a global functional equation.

7.2 Regularity gates for Cauchy-type equations

See [8, 9] for classical proofs that the gates below exclude Hamel-basis solutions.

Gate	**Typical justification**	**Use**
Borel measurable	$H' = \operatorname{deriv} H$ (or any measurable model)	excludes Hamel solutions
Monotone on an interval	order/convexity assumptions	implies measurability
Continuous at one point	smoothing / convolution heuristics	implies measurability

Table 1: Standard "regularity gates" that make Cauchy's equation rigid.

Route	**Gate G (Globality)**	**Gate R (Regularity)**	**Result**
Variational on H'	Universality $\to$ global equation	Regularity excludes Hamel	$H'(m) = B + C\log m$
KL atom $g(q) = d(1, q)$	System-independence $\to$ Cauchy	Measurability $\to g(q) = C\log q$	$d(p, q) = Cp\log(p/q)$
Entropy axioms	Axioms quantify over all distributions	Continuity excludes exotics	$H(p) = -K\sum p_i \log p_i$

Table 2: Entropy/KL derivations share two hypothesis gates: globality (Gate G) and regularity (Gate R).

Remark 7.2 (What Gate R excludes—necessity witness). Without a regularity gate, Cauchy's equation $f(x + y) = f(x) + f(y)$ admits uncountably many solutions. **Example (pathological solutions):** Let $\mathcal{B}$ be a Hamel basis for $\mathbb{R}$ over $\mathbb{Q}$. Define $f(x)$ by expressing $x = \sum_i q_i b_i$ (finite sum, $q_i \in \mathbb{Q}$, $b_i \in \mathcal{B}$) and setting $f(x) = \sum_i q_i g(b_i)$ for any function $g : \mathcal{B} \to \mathbb{R}$. Every such f is additive, but if g is chosen erratically, f is nowhere continuous, unbounded on every interval, and non-measurable. Any regularity gate kills these pathologies.

7.3 Solving the variational functional equation

The classification step is a standard rigidity phenomenon for Cauchy-type functional equations; see [8, 9].

Theorem 7.3 (Variational representation)**.** ***Assumptions:***

- $H' : (0, \infty) \to \mathbb{R}$ *satisfies the separated functional equation* $H'(m_x m_y) = \lambda(m_x) + \mu(m_y)$ *for all* $m_x, m_y > 0$.
- *Gate G (Globality): the equation holds globally, not just at specific stationary points.*
- *Gate R (Regularity):* H' *is Borel measurable (or monotone, or continuous at one point).*

Conclusion: *There exist constants* $B, C \in \mathbb{R}$ *such that*

$$H'(m) = B + C \log m \qquad (m > 0).$$

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

Integrating yields the entropy/divergence normal form

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

7.4 From the normal form to KL divergence and Shannon entropy

Specializing H to a divergence between two nonnegative vectors w, u with equal total mass gives the Kullback–Leibler expression

$$D(w\|u) = \sum_i w_i \log \frac{w_i}{u_i}$$

in the absolutely continuous case; in extended form one sets $D(w\|u) = +\infty$ if some $w_i > 0$ but $u_i = 0$.

For a probability distribution p on n states, Shannon entropy appears as the special case

$$S(p) = -\sum_{i=1}^{n} p_i \log p_i = \log n - D(p \,\|\, \text{uniform}).$$

Entropy is the KL divergence from p to the uniform distribution.

Countable/discrete measures. The same formulas make sense for countable state spaces as series, and they can be identified with the standard measure-theoretic KL divergence klDiv used in modern probability texts.

7.5 A clean comparison point: Shore–Johnson (1980)

Shore & Johnson propose a different "universality across applications" idea: a set of four consistency axioms for inference procedures that update prior distributions given constraints.

1. **Uniqueness.** The procedure should yield a unique result.
2. **Coordinate invariance.** The result should not depend on coordinate system choice.
3. **System independence.** For product systems $A \times B$, updating on constraints that factor as $C_A \times C_B$ should give the same result as updating each system separately.
4. **Subset independence.** If a constraint already determines some coordinates, the procedure on remaining coordinates should be unaffected.

The mathematical core of their argument runs as follows. Suppose the "atomic divergence" $d(p, q)$ contributes to a total divergence $D(P\|Q) = \sum_i d(p_i, q_i)$ that is minimized by inference procedures. System independence (axiom 3) forces d to satisfy an additivity identity over product distributions:

$$\sum_{i,j} d(p_i r_j,\, q_i s_j) \;=\; \sum_i d(p_i, q_i) + \sum_j d(r_j, s_j).$$

Testing this identity on Dirac delta distributions (concentrating all mass on single coordinates) extracts a multiplicative functional equation for $g(q) := d(1, q)$:

$$g(q_1 q_2) = g(q_1) + g(q_2), \qquad 0 < q_1, q_2 \leq 1.$$

This is precisely the multiplicative Cauchy equation, whose solutions are either $g(q) = C \log q$ (the "regular" solutions) or pathological Hamel-basis constructions (everywhere discontinuous, non-measurable). A regularity hypothesis—such as Borel measurability of g—excludes the pathological solutions and yields the logarithmic form $d(1, q) = C \log q$. Within the class of ratio-form atoms $d(p, q) = p \cdot g(p/q)$, this forces

$$d(p, q) = C\, p \log(p/q),$$

the KL divergence atom (up to a multiplicative constant).

The structural parallel with Knuth–Skilling is notable: both derivations reduce the problem to Cauchy-type functional equations, and both require an explicit regularity gate (measurability, monotonicity, or continuity) to exclude pathological solutions. The algebraic route arrives at addition and multiplication on the scale via lattice symmetries; Shore–Johnson arrives at the same functional forms via consistency requirements on inference procedures.

Remark 7.4 (On the strength of "system independence"). Uffink [20] shows that weaker formulations of "system independence" admit a broader Rényi family, not just KL. Our formalization bypasses this by explicitly requiring *product-additivity* of the divergence (the displayed equation above), which is the strengthening that singles out KL. Together with a regularity gate on g, this forms the explicit assumption ledger for the Shore–Johnson route.

8 Extension to σ-additivity

A standard limitation of axiomatic approaches—including Cox–Jaynes—is that they derive only finite additivity, while Kolmogorov's measure-theoretic foundations include σ-additivity from the start.[7] We show that this limitation is addressed by natural extension axioms.

8.1 The completeness axioms

The extension to countably infinite cases requires three additional axioms:

Axiom 8.1 (Sigma-complete events)**.** The event algebra admits countable joins: for any sequence (e_n) of events, the supremum $\bigsqcup_n e_n$ exists.

Axiom 8.2 (Scale completeness)**.** The scale S is sequentially complete: every bounded monotone sequence (s_n) has a supremum.

[7] This is a longstanding foundational debate. De Finetti [17] argued that finite additivity suffices and that σ-additivity is overly restrictive; modern Bayesians remain divided on whether σ-additivity is philosophically necessary or merely mathematically convenient.

Axiom 8.3 (Scott continuity)**.** The valuation $v : E \to S$ is Scott continuous: for any directed family[8] $(e_i)_{i\in I}$, $v(\bigsqcup_i e_i) = \bigsqcup_i v(e_i)$.

These axioms are **necessary**—they cannot be derived from the basic K&S axioms—but they are mathematically natural extensions that preserve the core algebraic structure. The Archimedean property (no infinitesimals) already constrains the scale to embed into $\mathbb{R}$, and scale completeness is the natural additional requirement for handling infinite operations.

Remark 8.4 (Categorical perspective: σ-frames)*.* From a categorical viewpoint, moving from Boolean algebras to σ-algebras corresponds to moving from frames to *σ-frames* (lattices with countable joins and finite meets, where finite meets distribute over countable joins). The Scott continuity axiom is the natural morphism condition in this setting. This connects the algebraic foundations to pointless topology and locale theory, where σ-additivity becomes structural rather than axiomatic.

8.2 The σ-additivity theorem

Theorem 8.5 (σ-additivity from K&S + completeness)**.** *Under the K&S axioms plus the three completeness axioms above, the composition $\mu = \Theta \circ v$ (where $\Theta : S \to \mathbb{R}$ is the additive representation) is σ-additive for pairwise disjoint sequences:*

$$\mu\Big(\bigsqcup_n e_n\Big) = \sum_{n=0}^{\infty} \mu(e_n).$$

8.3 Significance

The key insight is that $\oplus \to +$ is **derived** from the K&S axioms (via the representation theorem), not assumed. The additive structure that connects to standard measure theory comes from the K&S derivation, not from an external assumption.

This resolves the "finite-only" criticism of algebraic foundations: *K&S naturally extends to σ-algebras when appropriate completeness axioms are added, and the extension is grounded in the same algebraic principles that govern the finite case.*

9 How this compares to classical foundations

The following table summarizes the key structural assumptions of each foundational approach:

Foundation	Events	Codomain	Key axiom	Regularity gate
Kolmogorov	σ-algebra	$[0, 1]$	σ-additivity	—
Cox–Jaynes	propositions	$[0, 1]$	consistency	continuity
de Finetti	events	$[0, 1]$	finite add.	—
K&S (finite)	lattice	ordered scale	associativity	no anomalous pairs
K&S + compl.	σ-lattice	complete scale	+ Scott cont.	+ completeness
Shore–Johnson	distributions	divergence	system indep.	measurability

Each foundation makes different choices about what to assume and what to derive:

- **Kolmogorov** postulates σ-additivity directly: $P : \mathcal{F} \to [0, 1]$ with $P(\Omega) = 1$ and countable additivity on disjoint unions. [3]

[8] A family where any two elements have an upper bound in the family—the natural generalization of "increasing sequence" to non-linear index sets.

- **Cox–Jaynes** derives the sum and product rules from consistency requirements on plausibility assignments, given a continuity hypothesis. [2]
- **De Finetti** takes finite additivity as primitive, arguing that σ-additivity is an unnecessary strengthening. [17]
- **K&S** derives additivity and multiplication from lattice symmetries (associativity, monotonicity), given a "no anomalous pairs" hypothesis on the scale.
- **Shannon–Faddeev** derives entropy uniqueness from axioms on probability vectors (continuity, maximality at uniform, additivity over independent systems). [5, 7]
- **Shore–Johnson** derives KL divergence from consistency axioms on inference procedures, given a measurability hypothesis. [4]

In this sense, the algebraic program and algorithmic probability (Solomonoff induction) are complementary: the former derives the *form* of Bayesian updating from symmetry axioms, the latter derives a canonical *prior* from computability and universality constraints [18, 19]. Our results apply equally to such algorithmic priors once they are normalized to probability measures.

The approaches that *derive* the calculus (Cox, K&S, Shore–Johnson) all face the same mathematical issue: functional equations admit pathological Hamel-basis solutions unless a regularity hypothesis excludes them. A careful foundation keeps this gate visible.

Remark 9.1 (Where the novelty sits). The novelty of the algebraic approach is not the final calculus (which agrees with Kolmogorov/Shannon) but the *upstream* identification of which symmetry principles suffice to force the calculus, and which regularity principles are logically indispensable.

Remark 9.2 (Imprecise probabilities). Not all applications justify a single sharp measure. Representing uncertainty by sets of measures (credal sets) or interval bounds corresponds algebraically to weakening the scale from a total order to a partial order.[9] Conversely, our formalization proves that total order is *necessary* for point-valued representations: if the scale has incomparable elements, no faithful embedding into $\mathbb{R}$ exists.

A Dictionary: Paper notation to Lean names

The following table maps the mathematical concepts in this paper to their Lean formalizations in the `Mettapedia.ProbabilityTheory.KnuthSkilling` namespace.

[9] See Walley [16].

Paper term	Lean name	File
Ordered scale $(S, \oplus)$	`KSSemigroupBase`	`Core/Basic.lean`
$+$ identity as minimum	`KnuthSkillingAlgebraBase`	`Core/Basic.lean`
Separation axiom	`KSSeparation`	`Core/Algebra.lean`
No anomalous pairs	`NoAnomalousPairs`	`Additive/Axioms/AnomalousPairs.lean`
Hölder embedding	`holder_embedding`	`Additive/Proofs/OrderedSemigroupEmbedding/HolderEmbedding.lean`
$\oplus$, $\otimes$, $\star$	`op`, `tensor`, `condOp`	`Core/Basic.lean`
$x^{(n)}$ (iteration)	`iterate_op_pnat`	`Core/Basic.lean`
Additive representation	`representation_theorem`	`Additive/Main.lean`
Product representation	`product_representation`	`Multiplicative/Main.lean`
Variational theorem	`variational_classification`	`Variational/Main.lean`
KL divergence	`klDiv`	`Information/DivergenceMathlib.lean`
Regraduation	order isomorphism	various
Sigma-complete events	`SigmaCompleteEvents`	`Core/ScaleCompleteness.lean`
Scale completeness	`KSScaleComplete`	`Core/ScaleCompleteness.lean`
Scott continuity	`KSScottContinuous`	`Core/ScaleCompleteness.lean`
σ-additivity theorem	`ks_sigma_additive`	`Core/ScaleCompleteness.lean`
$\oplus \to +$ (derived)	`op_is_addition_via_Theta`	`Additive/Axioms/OpIsAddition.lean`
Mathlib bridge	`toProbabilityMeasure`	`Bridges/MathlibProbability.lean`

References

[1] Kevin H. Knuth and John Skilling. *Foundations of Inference*. Axioms, 1(1):38–73, 2012.

[2] Richard T. Cox. *Probability, Frequency and Reasonable Expectation*. American Journal of Physics, 14(1):1–13, 1946.

[3] Andrey N. Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer, 1933.

[4] J. E. Shore and R. W. Johnson. *Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-Entropy*. IEEE Trans. Inform. Theory, 26(1):26–37, 1980.

[5] C. E. Shannon. *A Mathematical Theory of Communication*. Bell System Technical Journal, 27:379–423 and 623–656, 1948.

[6] A. I. Khinchin. *Mathematical Foundations of Information Theory*. Dover, 1957.

[7] D. K. Faddeev. *On the concept of entropy of a finite probabilistic scheme*. Uspekhi Mat. Nauk, 11(1):227–231, 1956.

[8] J. Aczél. *Lectures on Functional Equations and Their Applications*. Academic Press, New York, 1966.

[9] M. Kuczma. *An Introduction to the Theory of Functional Equations and Inequalities: Cauchy's Equation and Jensen's Inequality*. 2nd ed., edited by A. Gilányi. Birkhäuser, Basel, 2009.

[10] N. G. Alimov. On ordered semigroups. *Izv. Akad. Nauk SSSR Ser. Mat.*, 14:569–576, 1950.

[11] L. Fuchs. *Partially Ordered Algebraic Systems*. Pergamon Press, 1963.

[12] O. Hölder. Die Axiome der Quantität und die Lehre vom Mass. *Berichte über die Verhandlungen der Königlich Sächsischen Gesellschaft der Wissenschaften zu Leipzig, Mathematisch-Physische Classe*, 53:1–64, 1901.

[13] R. T. Rockafellar and R. J.-B. Wets. *Variational Analysis*. Springer, 1998.

[14] F. H. Clarke. *Optimization and Nonsmooth Analysis*. SIAM, 1990.

[15] Eric Luap. *OrderedSemigroups: Formalization in Lean 4*. GitHub repository, 2024. `https://github.com/ericluap/OrderedSemigroups`

[16] Peter Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman & Hall, 1991.

[17] Bruno de Finetti. *Theory of Probability* (2 vols). Wiley, 1974–1975.

[18] Ray J. Solomonoff. *A Formal Theory of Inductive Inference. Part I and Part II*. Information and Control, 7(1):1–22 and 7(2):224–254, 1964.

[19] Marcus Hutter. *Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability*. Springer, 2005.

[20] Jos Uffink. *Can the Maximum Entropy Principle Be Explained as a Consistency Requirement?* Studies in History and Philosophy of Modern Physics, 26(3):223–261, 1995.