Relative Probability on Finite Sample Spaces
SUBTITLE HERE

Max Sklar
Local Maximum Labs
DATE HERE

Abstract

This is an incomplete draft/outline of an upcoming paper. Please do not share

Contents

1 Introduction

The foundations of probability theory are still very much open to debate!

Since Kolmogorov published the standard axioms for probability[9] in 1933, there have been calls to relax or alter them for various applications. In Kolmogorov's Axiomatisation and Its Discontents[5], Lyon lays out these cases and their justifications. One controversy concerns conditional probability. We talk about "the probability of A given that B occured" and denote it as $P(A|B)$ - but can this make sense if B never occurs (or has probability zero[1])? We are out of luck if we stick to the Kolmogorov model, which defines the conditional probability above as the ratio $\frac{P(A\cap B)}{P(B)}$. If the probability of B is zero, the indeterminate form $\frac{0}{0}$ appears.

Undeterred, mathematicians and engineers refer to this type relative probability all the time. For example, if we consider a continuous probability distributrion over $[0, 1]$ given by the probability distribution function $2x$, we know that the PDF at $x = \frac{1}{2}$ is twice as much as the PDF at $x = \frac{1}{4}$. In a sense, we believe that the former is twice as likely as the latter - even though we are only talking about *probability density*. Hajek[6] (citing Borel) gives a much more compelling example: if a random point on the Earth is selected, what is the probability that it is in the eastern hemisphere given that it is on the equator? Most people would not hesitate to answer one half, and yet the equator - being a mere 1-dimentional object - has probability 0 compared to the rest of the globe.

Let us take the position that we may model probability in a non-standard way, and we can do so as long our new framework is logically consistent, and the advertised applications correspond to the new model[2]. We ought to understand whether we can derive a framework for probability that takes the relationships between outcomes and events as the fundamental unit.

This improves the Kolmogorov model - that starts with an absolute probability function - by both solving the conditional probability question and giving rise to new objects and patterns to study.

1.1 Goals

The relative approach to probability is not only viable, but has many properties that practicioners will find attractive. For example, many Bayesian inference algorithms rely on relative probability alone to search for optimal parameters.

As a proof of this concept, we will construct a theory of relative probability on finite distributions. By omitting infinite distributions, we temporarily set aside the concepts of measurable sets and countable additivity[3]. This work will demonstrate that even with this vast simplification there is much to be learned. Relative probability requires a new set of fundamental rules and vocabulary, which we will construct without the distractions of infinite and continuous outcome spaces.

[1] Another unintuitive feature of probability theory is that zero probability events do indeed occur, particularly when given a continuous distribution.

[2] Lyon identifies this link between application and model as the bridge principle. A new set of axioms for probability could well give rise to a new and interesting mathematics, but if that mathematics cannot be linked to any application that anyone would reasonbly call probability, then it ought to go by a different name.

[3] In the textbook Invitation to Discrete Mathematics, Matoušek et al. write

> By restricting ourselves to finite probability spaces we have simplified the situation considerably... A true probability theorist would probability say that we have excluded everything interesting.

We then provide a new formulation of Bayes rule that uses only relative proability, and that is reflective of current practice. We will look at these applications of these ideas along with their algorithmic implementation.

Finally, we discuss a critical feature of relative probability functions, which is their ability to retain information when taking limits—for which absolute probability fails. To that end, we delve into the topology of the relative probability space, and finish with a proof of compactness.

2 Preliminaries

2.1 Magnitude Space

Definition 2.1. The *magnitude space* $\mathbb{M}$ is the set of all positive real numbers, 0 and ∞.

$$\mathbb{M} = [0, +\infty]$$

Magnitudes roughly corresponding with our intuition of size. Unlike the set of non-negative real numbers, magnitudes include a point at infinity. The point at infinity can be considered a limit element, larger that all of the other magnitudes. It provides the magnitude space with several important properties.

1. *Compactness* : sequences that go off to infinity could still be considered to have a limit point at ∞
2. Symmetry around ratios: When we compare the probability of two events, we get their *odds*. If the odds are 0, then we are comparing an event with probability 0 to an event with probability ¿ 0. We should be able to reverse this comparison, and talk about it from either side. If we're comparing the probability of an event to its converse, then ∞ represents events that have probability 1.
3. Actual measurement value: Whereas physical objects may not be able to have infinite size, mathematical objects do. The infinite element is introduced in measure theory because many mathematical systems (real numbers for one) necessarily contain subsets of infinite measurement.

We will set $0^{-1} := \infty$ and $\infty^{-1} := 0$, even though their product is indeterminate within $\mathbb{M}$ and they do quite act as a multiplicative inverses of each other.

2.2 The Wildcard Element

Definition 2.2. Let the *magnitude-wildcard space* $\mathbb{M}^* = \mathbb{M} \cup \{*\}$ be the set of magnitudes along with a *wildcard element*, $*$.

The wildcard element corresponds to several different concepts, each appearing in a different type of practice:

- The *NaN*, or *Not a Number*[4] value in the IEEE standard for floating point arithmetic[8].
- The indeterminate form $\frac{0}{0}$ in arithmetic.

[4] "Not a Number" may have been an unfortunate naming choice because it actually represents **any** number!

- The *wildcard pattern* used in pattern matching and regular expressions in type theory and computer science

The following properties on $*$ to allow addition and multiplication of any two magnitude-wildcard values.

(i) $0 \cdot \infty = *$

(ii) $* + m = *$

(iii) $* \cdot m = *$

Note that we now lose some basic properties of these operations. For example, we can no longer simplfy an expression like $0x$ to 0. This will take some getting used to, but programmers familiar with the floating point value *NaN* have long adapted to this.

2.3 The Matching Relation

Definition 2.3. The *matching relation*[5] $:\cong$ is a binary relation on $\mathbb{M}^*$. m_1 is matched by m_2 when either $m_1 = m_2$ or m_2 is the wildcard.

$$m_1 :\cong m_2 \iff (m_1 = m_1) \vee (m_2 = *)$$

The left hand side of a matching relation is the *parameter* and the right hand side is the *constraint*. This reinforces the idea that the *contraint* may or may not constrain the parameter. The wildcard element represents every single value, but it cannot be represented by any specific value. Alternatively, the wildcard element also represents a loss of information about a parameter which can never be recovered.

The following lemmas quickly follow from the definition.

Lemma 2.1. *If a magnitude matches a non-wildcard element, then the two values are equal.*

$$m_1 :\cong m_2 \wedge m_2 \neq * \implies m_1 = m_2$$

Lemma 2.2. *Every element is matched by the wildcard element.* $m :\cong *$

Lemma 2.3. *The wildcard element is matched only by itself.* $* :\cong m \implies m = *$

The matching relation looks a lot like equality, and in many cases it is, but because of the introduction of the wildcard it doesn't always act in the same way.

Theorem 2.4. *The matching relation is reflexive and transitive, but unlike equality is not symmetric.*

Proof. Reflexive is obvious: $m :\cong m \iff (m = m) \vee (m = *)$

The transitive property states that for all m_1, m_2, m_3 in $\mathbb{M}$, if $m_1 :\cong m_2$ and $m_2 :\cong m_3$, then $m_1 :\cong m_3$.

Assume that $m_1 :\cong m_2$ and $m_2 :\cong m_3$. If none of these values are the wildcards, then by property 2.1, they are all equal and $m_1 :\cong m_3$. If $m_1 = *$ then by property 2.3, $m_2 = *$ and finally $m_3 = *$ so the theorem

[5] It helps to read $:\cong$ as "is matched by".

holds. If $m_2 = *$ then $m_3 = *$ and $m_1 :\cong m_3$ by property 2.2. And of course if $m_3 = *$ alone, then by property 2.2, m_1 is still matched by m_3.

For non-symmetric, we present a counterexample: $1 :\ncong * \text{ but } * :\ncong 1$ □

We could also define a symmetric matching relation $m_1 :\cong: m_2$ to mean $m_1 :\cong m_2 \vee m_2 :\cong m_1$. This would be symmetric, but not transitive.

Theorem 2.5. *The matching relation preserves multiplication and addition.* $\forall a, b, a', b' \in \mathbb{M}^*$ *if* $a :\cong a'$ *and* $b :\cong b'$*, then* $ab :\cong a'b'$ *and* $a + b :\cong a' + b'$*.*

Proof. For multiplication: Let $a, b, a', b' \in \mathbb{M}^*$, and let $a :\cong a'$ and $b :\cong b'$. If either a' or b' are wildcards, then $a'b'$ is also a wildcard. If a' and b' are not wildcards, then $a = a'$ and $b = b'$, also making $ab :\cong a'b'$. The same argument proves $a + b :\cong a' + b'$. □

3 Categorical Distribution

Let Ω be a set of mutually exclusive *outcomes*[6]. We assume that Ω is finite so that we can count its members as $|\Omega| = K$. We say there are K outcomes, or *categories*.

Definition 3.1. A *categorical distribution* on a Ω is a function $P : \Omega \to [0, 1]$ such that $\sum_{h\in\Omega} P(h) = 1$

The set of all categorical distributions of size K can be embedded in $\mathbb{R}^K$ as a (K-1)-dimentional object called a (K-1)-simplex (see figure 1). For example, if $K = 3$, the resulting space of categorical distributions is an equilateral triangle embedded in $\mathbb{R}^3$ connecting the points (1, 0, 0), (0, 1, 0), and (0, 0, 1).

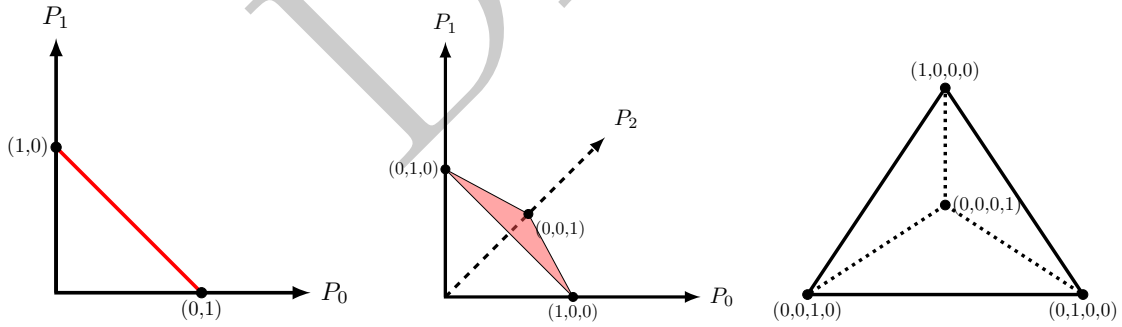


Figure 1: An illustration of probability simplecies for K = 2, 3, and 4. These objects are respectively, a segment embedded in $\mathbb{R}^2$, an equilateral triangle embedded in $\mathbb{R}^3$, and a normal tetrahedron embedded in $\mathbb{R}^4$. We make no attempt to visualize the 4D space that contains the tetrahedron.

With absolute probability, information about relative probability is lost at the vertices where several probabilities might go to zero. For example, if $\Omega = a, b, c$ with $P(a) = 1$ and $P(b) = P(c) = 0$, we cannot compare the probabilities of b and c as we can in the rest of the simplex.

[6]Each outcome could be thought of as a possible result of a random trial, or a possible outcome for an unknown variable

This poses an interesting problem for limits. Consider the following categorical distribution function, with parameter $\epsilon > 0$:

$$P(a) = 1 - \epsilon P(b) = \frac{2}{3}\epsilon P(c) = \frac{1}{3}\epsilon$$

This is clearly an absolute probability, and its clear that the limit as ϵ goes to zero should be $P(a) = 1, P(b) = P(c) = 0$. The fact that b is twice as likely as c is lost!

3.1 Events

An *event* is a set of outcomes, and by convention $\mathcal{F}$ is the set of all possible events. $\mathcal{F}$ is the power set[7] of Ω, meaning that $\mathcal{F} = \mathcal{P}(\Omega)$, and for any subset $e \subseteq \Omega$, $e \in \mathcal{F}$.

In the previous section, we defined the probability of individual outcomes. We can now define the probability of an event - that is the probability that any one of its outcomes occur. Looking at probability on the event level rather than the outcome level is a crucial insight in the development of probability theory (and measure theory more generally). Even though the process is far simpler for finite distributions, we must pay attention to this layer in order for the framework to generalize.

For all e in $\mathcal{F}$,

$$P(e) = \sum_{h \in e} P(h)$$

We can take P as acting either on outcomes or events using the convention $P(\{h\}) = P(h)$.

Ω itself the *universal event* of all outcomes, with probability 1.

$$P(\Omega) = \sum_{h \in \Omega} P(h) = 1$$

3.2 Relative Probability Function

A *relative probability function*, or *RPF*, measures the probability of one event with respect to another. For example, we may wish to talk about an event that is "twice as likely" as another, even if we don't know the absolute probability of either event.

We continue to use P to represent the RPF but with two inputs instead of one. The expression $P(e_1, e_2)$ can be read as the probability of e_1 relative to e_2.

$$P : \mathcal{F} \times \mathcal{F} \to \mathbb{M}^*$$

We define relative probability in terms of absolute probability as a ratio, in the style of the standard Kolmogorov framework.

[7] In general, $\mathcal{F}$ is not the entire power set of Ω but typically is when Ω is finite. We need not concern ourselves with the σ-algebra of measurable sets here.

Definition 3.2. The relative probability of events e_1 and e_2 on an categorical distribution P is given as

$$P(e_1, e_2) = \frac{P(e_1)}{P(e_2)}$$

If $P(e_1) = P(e_2) = 0$, then $P(e_1, e_2) = *$, representing the classical problem of zero-probability events being incomparable.

Theorem 3.1 (Composition)**.** *For all events e_1, e_2, e_3, $P(e_1, e_3) :\cong P(e_1, e_2) \cdot P(e_2, e_3)$*

Proof. Start with the case that $P(e_2) \neq 0$. Then $P(e_1, e_2) \cdot P(e_2, e_3) = \frac{P(e_1)}{P(e_2)} \frac{P(e_2)}{P(e_3)} = \frac{P(e_1)}{P(e_3)} = P(e_1, e_3)$. When $P(e_2) = 0$, $P(e_1, e_2) \cdot P(e_2, e_3) = \frac{P(e_1)}{P(e_2)} \frac{P(e_2)}{P(e_3)} = *$. Because $*$ matches everything, then the matching statement holds. Because it holds in both cases, the theorem is true. □

4 The Relative Probability Approach

In section 3.2, the relative probability function was derived from the absolute probability function. Here in section 4, we start with the relative probability function as the fundamental object of study.

4.1 Fundamental Axioms

Consider a relative probability function P that acts on outcomes only.

Definition 4.1. Let Ω be the set of outcomes, and $P : \Omega \times \Omega \to \mathbb{M}^*$ be a function acting on two outcomes to produce a magnitude-wildcard. P is a *relative probability function on the outcomes of* Ω if it obeys the *3 fundamental axioms of relative probability*:

(i) The *identity axiom*: $P(h, h) = 1$

(ii) The *inverse axiom*: $P(h_1, h_2) = P(h_2, h_1)^{-1}$

(iii) The *composition axiom*: $P(h_1, h_3) :\cong P(h_1, h_2) \cdot P(h_2, h_3)$

If P is a relative probability function, $P(h_1, h_2)$ can be read as the probability of h_1 relative to h_2. Outcomes h_1 and h_2 are said to be *comparable* if $P(h_1, h_2) \neq *$.

Let us pause for a moment to discuss how these axioms where chosen. The star of the show is the composition axiom which succinctly encodes how relative probability works. If A is twice as likely as B, and B is 3 times as likely as C, then A had better be 6 times as likely as C. If it were not, then these relative probability assignments would have no meaning.

The composition axiom is enough to show that the identity axiom works most of the time. For example, if one can compare an outcome h_1 to any other outcome h_2 then through composition we get $P(h_1, h_2) :\cong P(h_1, h_1) \cdot P(h_1, h_2)$. So long as $P(h_1, h_2)$ isn't 0, ∞, or $*$, then we would have to conclude $P(h_1, h_1) = 1$.

But that doesn't get us all the way there! We can still construct scenarios where $P(h, h) = *$. Hence, the neccesity of the identity axiom.

Composition and identity can actually be combined into a single axiom about composition paths. It's a bit more unweildy for the mathematical proofs, but nevertheless interesting.

Proposition 4.1 (Path Composition)**.** *Given a non-empty list of N outcomes $h_0, h_1, h_2, ..., h_{N-1}$,*

$$P(h_0, h_{N-1}) :\cong \prod_{k=0}^{N-2} P(h_k, h_{k+1})$$

In this case, $P(h_0, h_0)$ would be matched by the empty product, which is 1.

The inverse axiom is nearly redundant as well. Since $P(h_0, h_0) \cong P(h_0, h_1) \cdot P(h_1, h_0)$, the terms in the constaint look like they must be inverses! But without stating the axiom explicitly, there could be a case where $P(h_0, h_1)$ is some non-wildcard magnitude like 2 but $P(h_1, h_0)$ is not comparable. This shouldn't be allowed because $*$ represents a lack of knowledge about a value, and we consider $P(h_1, h_0)$ and $P(h_1, h_0)$ to be the exact same piece of information but in reverse.

4.2 Examples

Definition 4.2. The *uniform* RPF can be constructed from any number of outcomes where each are considered equally likely. $P(h_1, h_2) = 1$ for every pair of outcomes.

Definition 4.3. The *uncomparable* RPF has $P(h_1, h_2) = *$ for every pair of outcomes.

Perhaps it is seens unpolished to put such a qualitative statment towards a math object, but the uncomparable RPF is a real downer! It's as if the observer gave up.

Definition 4.4. A *certain* RPF contains a single outcome that has infinite probability relative to all other outcomes. Let h_C be the certain outcome with $h_C \neq h$. Then $P(h_C, h) = \infty$. The relative probability of the other $K - 1$ outcomes could be anything.

Definition 4.5. The *empty* RPF has no outcomes $K = 0$, and therefore the function P has no valid inputs.

It is surprising that there is still an RPF with $\Omega = \varnothing$.This is an interesting comparison to absolute distributions where such a function does not exist (because with no outcomes, they cannot sum to 1).

Definition 4.6. The *unit* RPF has a single outcome where $K = 1$ and $\Omega = h$. There is only one such RPF where $P(h, h) = 1$.

The unit RPF is a special case of the uniform RPF and the certain RPF. This matches the absolute case where the probability of the single outcome must be 1.

Definition 4.7. Let P be an RPF with K outcomes labeled $(h_0, h_1, ..., h_{K-1})$. P is a *finite geometric* RPF with ratio r if the relative probabilities of each outcome with its neighbor is always r. In other words, for all $i \in (0, 1, ..., K - 2)$,

$$P(h_{i+1}, h_i) = r$$

When r is 0 or ∞, we can call this the *limit finite geometric* RPF.

Finally, to include an example that is both common and has powerful applications, there is a relative version of the Binomial distribution.

Definition 4.8. A *binomial distribution* has a sample size we can call n, and a probability of success p. The RPF is uses$\Omega = \{0, 1, 2, ..., n\}$, and thus $K = n + 1$. It is given as follows:

$$P(h_1, h_2) = \frac{h_2!(n - h_2)!}{h_1!(n - h_1)!} \left(\frac{p}{1 - p}\right)^{h_1 - h_2}$$

5 New Concepts for Relative Probability

We have successfully defined the relative probability in section 4 with fundamental axioms and have constructed some examples. Because new situations arise that do not occur in the Kolmogorov model, we need to define some new vocabulary.

Fortunately, we can look at the absolute probability function as a special case of relative probability, defined by $P(h_1, h_2) = \frac{P(h_1)}{P(h_2)}$.

Figure 2 gives us a roadmap of these new concepts and their relationship to each other.

5.1 Matching and Comparability

Definition 5.1. A relative probability function is *totally comparable* if every pair of outcomes are comparable.

Theorem 5.1. *An absolute probability function is totally comparable if and only if $P(h) = 0$ for at most one outcome.*

Proof. Let P be an **absolute** probability function, with h_1 and h_2 being two outcomes. If $P(h_1) = P(h_2) = 0$, then $P(h_1, h_2) = \frac{0}{0} = *$. If only outcome h_1 is assigned 0, then $P(h_1, h_1) = 1$, $P(h_1, h_2) = 0$, and $P(h_2, h_1) = \infty$. Any other pairing that does not involve h_1 will be the quotient of two positive numbers, and thus also comparable. □

Definition 5.2. An *anchored* RPF has at least 1 outcome whose probability relative to every other outcome is greater than zero. We call this outcome an *anchor* element, and there may be multiple.

The anchored concept is useful because it means that every pair of outcomes, even if they are not comparable, are at least going to be 0 compared to a larger outcome. The anchoring of a distribution is important to ensure that it is well behaved.

Theorem 5.2. *All absolute probability distributions are anchored.*

Proof. Let P be an absolute probability distribution on Ω. Because $\sum_{h \in \Omega} P(h) = 1$, there must be at least one h such that $P(h) > 0$. Therefore, for any comparison outcome h', $P(h, h') > 0$ □

Lemma 5.3. *Every non-empty, totally comparable RPF is anchored.*

Proof. Let P be non-empty and totally comparable RPF. Assume the opposite - that is for every outcome h, there exists another outcome h' such that P(h, h') = 0.

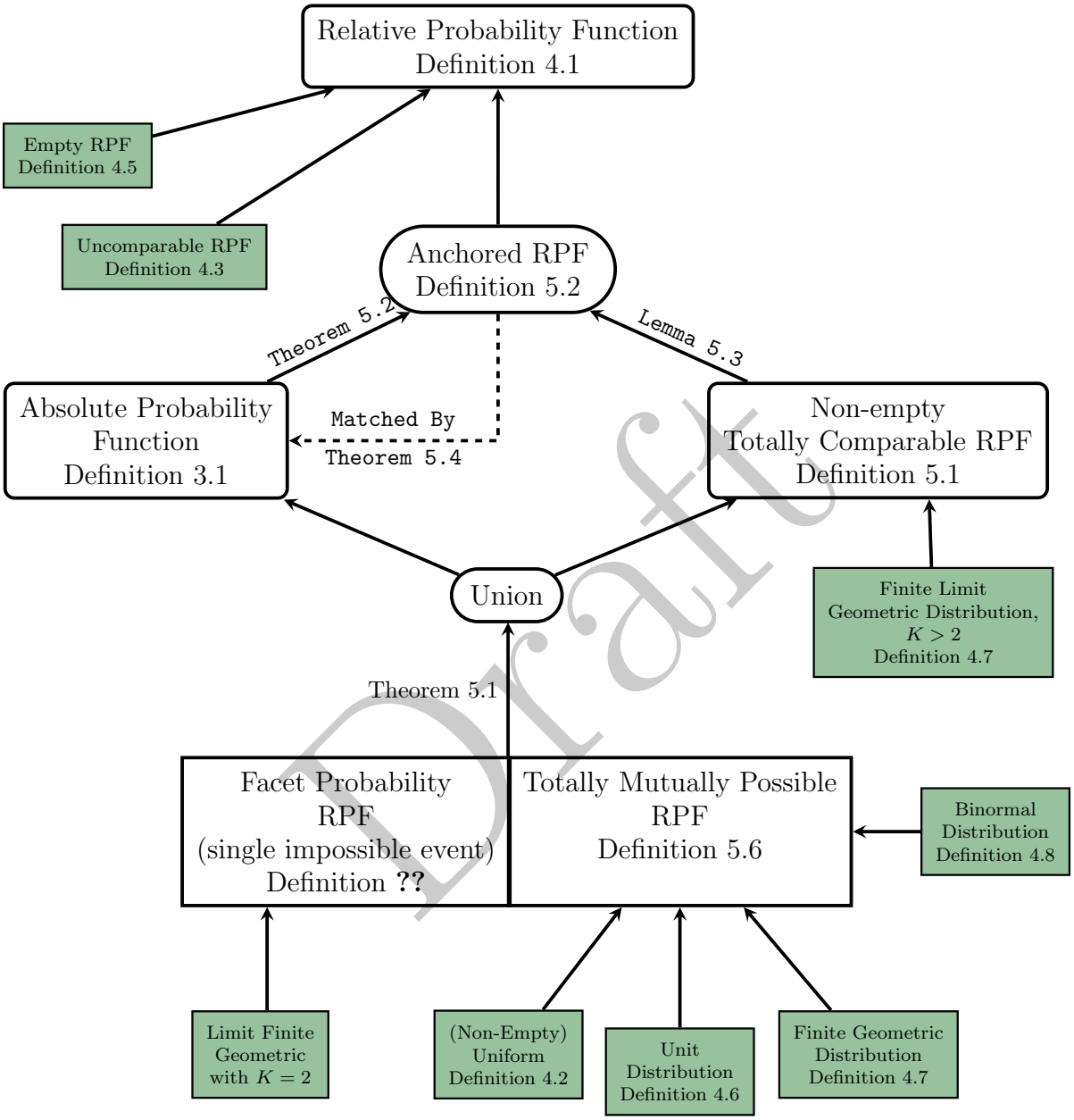


Figure 2: This is our roadmap for all of the sub-types of relative probability functions and their relationship to one another.

Therefore, a function $f : \Omega \to \Omega$ can be created so that for every h, $P(h, f(h)) = 0$.

Let $f^n(h)$ be the function f applied to h n times. Then $P(h, f^n(h)) = 0$ for all n greater than 0. This is by induction because the case of $n = 1$ was assumed above, and for for inductive step

$$P(h, f^{n+1}(h) :\cong P(h, f^n(h)) \cdot P(f^n(h), f(f^n(h)) = 0 \cdot 0 = 0$$

Because Ω is finite, repeated applications of f on h must evenually return to an outcome that has already been visited. In more rigorous terms, there exists an N such that $f^N(h) = f^i(h)$ for some $i < N$.

But this is a contradiction because $P(f^i(h), f^N(h))$ should equal 0 by the argument above, but 1 by the identity axiom. □

A totally comparable RPF contains the maximum amount of information about the relative probability of two events. Some RPFs may have less information but nevertheless are consistent with RPFs that have more. The following definition encapulates this relationship.

Definition 5.3. Let P_1 and P_2 be relative probability functions. P_1 is matched by P_2 if and only if all of relative probabilities of P_1 are matched by those of P_2.

$$\forall h_1, h_2 \in \Omega, P_1(h_1, h_2) :\cong P_2(h_1, h_2)$$

Theorem 5.4. *Every anchored RPF is matched by an absolute probability function, given by the following equation where a is an anchor outcome.*

$$P(h) = \frac{P(h, a)}{\sum_{h' \in \Omega} P(h', a)}$$

Proof. We need to show that $P(h)$ is a valid absolute probability function, and that it matches the original RPF.

Because a is an anchor element, we know that $P(h', a) < \infty$. This means that the sum $\sum_{h' \in \Omega} P(h', a) < \infty$. It is also non-zero, because included in that sum is $P(a, a) = 1$. The numerator $P(h, a)$ is also a magnitude $< \infty$. Therefore, this formula yields $P(h) \notin \{\infty, *\}$.

We next check that the values of $P(h)$ sum to 1 as follows:

$$\sum_{h \in \Omega} P(h) = \sum_{h \in \Omega} \frac{P(h, a)}{\sum_{h' \in \Omega} P(h', a)} = \frac{\sum_{h \in \Omega} P(h, a)}{\sum_{h' \in \Omega} P(h', a)} = 1$$

Cancellation of these equal sums is justified because we have argued above that they cannot be 0 or ∞.

Therefore, $P(h)$ is a valid absolute probability function. To show that relative probability function is matched by it:

$$P(h_1, h_2) :\cong P(h_1, a) \cdot P(a, h_2) = \frac{P(h_1, a)}{\sum_{h' \in \Omega} P(h', a)} \div \frac{P(h_2, a)}{\sum_{h' \in \Omega} P(h', a)} = \frac{P(h_1)}{P(h_2)} \tag{1}$$

□

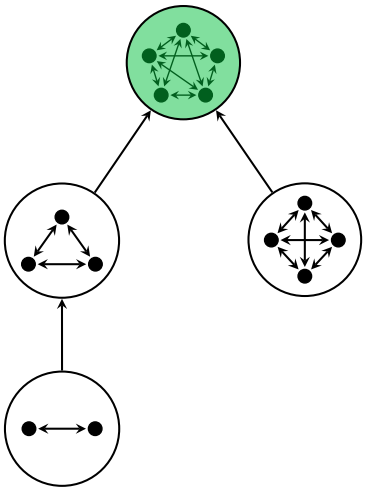


Figure 3: A diagram of an anchored RPF with its mutually possible classes. The anchor class is shaded.

5.2 Possibility Classes

Definition 5.4. Outcomes h_1 and h_2 are *mutually possible* if they are comparable and $0 < P(h_1, h_2) < \infty$.

Theorem 5.5. *The relationship of mutually possible events is an equivalence relation, being reflexive, symmetric and transitive.*

Proof. For reflexive, $P(h_1, h_1) = 1$ by the identity axiom.

For symmetric, $P(h_1, h_2) = P(h_2, h_1)^{-1}$, which means that each can be in $\{0, \infty, *\}$ if and only if the other one is as well.

For transitive, we use the composition axiom which states that $P(h_1, h_3) :\cong P(h_1, h_2) \cdot P(h_2, h_3)$. If the last 2 values are positive real numbers, then their product is also a positive real number and equal to $P(h_1, h_3)$. □

Definition 5.5. Outcome h_1 is *impossible* with respect to h_2 if $P(h_1, h_2) = 0$. Outcome h_1 is *possible* with respect to e_2 if they are comparable and $P(h_1, h_2) > 0$

Theorem 5.6. *The relationship of being possible is a preorder, being both reflexive and transitive.*

Proof. It must be reflexive because $P(h, h) = 1$. If $P(h_1, h_2) > 0$ and $P(h_2, h_3) > 0$ then their product is also greater than zero, and by composition, equal to $P(h_1, h_3)$. Thus h_1 is also possible with respect to h_3. □

If we consider a possibility relationship with respect to the equivalence classes of mutually possibility, then we have a partial order.

Theorem 5.7. *If an RPF is totally comparable, then the equivalance classes of mutually possible outcomes are totally ordered. That is, each member of an equivalence class of outcomes is comparable to each member of another class with that comparison always being 0 or ∞.*

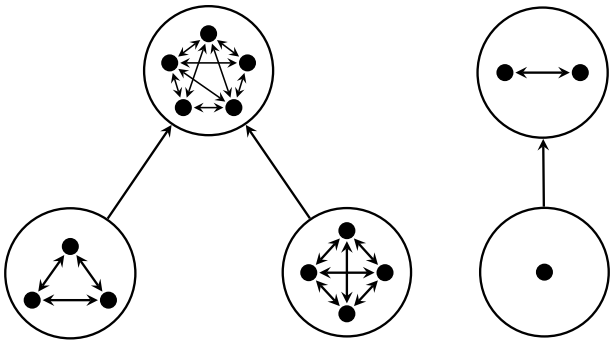


Figure 4: This is the diagram for a single RPF that is not anchored. Notice that there is no anchored mutually possible class. This meams that we cannot turn this into an absolute probability function.

Proof. Let A and B be 2 distinct mutually possible equivalence classes on Ω, and let $a \in A$ and $b \in B$. Then $P(a, b)$ must be either 0 or ∞ because if it were in between then a and b would be in the same equivalence class, and if it were $*$ then P wouldn't be totally comparable.

Let $a' \in A$ and $b' \in B$. Then $0 < P(a', a) < \infty$ and $0 < P(b, b') < \infty$ due to the definition of mutual comparability. Thus with composition we get

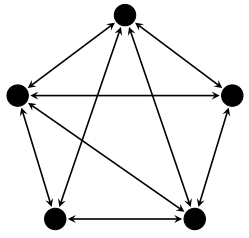
$$P(a', b') :\cong P(a', a) \cdot P(a, b) \cdot P(b, b') = P(a, b)$$

Therefore, all comparisons between the 2 classes will be the same, and they will either be 0 or ∞. □

5.3 Totally Mutually Possible RPFs

Definition 5.6. A relative probability function is called *totally mutually possible* if all of its outcomes[8] are mutually possible. Mutually possible RPFs satisfy *Cromwell's rule* in Bayesian inference, which states that prior beliefs should not assign probability zero or one to events[9].

Totally mutually possible RPFs have a simple diagram where all the outcomes are completely connected. All of the outcome are anchor outcomes, and there is a single mutually possible class.



Theorem 5.8. *A non-empty totally mutually possible RPF is equal to an absolute probability function.*

[8]Note that this one of the few definitions that cannot be upgraded from outcomes to events. The empty event $e = \{\}$ for example will be impossible with respect to any outcome by theorem 6.2.

[9]It would be stated differently in continuous space.

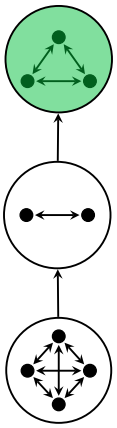


Figure 5: This is a diagram of a totally comparable RPF that is not mutually possible. The mutually possible components form a total order, with the *anchored component* - which contains all anchor elements - is on top.

Proof. If P is non-empty, and totally mutually possible, all of it's outcomes are anchors. Therefore, we can use theorem 5.4 to find a matching absolute probability function

$$P(h) = \frac{P(h,a)}{\sum_{h' \in \Omega} P(h',a)}$$

Because every element of Ω is an anchor, we can let $a = h$ and get

$$P(h) = \frac{P(h,h)}{\sum_{h' \in \Omega} P(h',h)} = \frac{1}{\sum_{h' \in \Omega} P(h',h)}$$

Theorem 5.4 states that $P(h_1, h_2) :\cong \frac{P(h_1)}{P(h_2)}$, but since the constraint is never $*$, they must be equal. □

We have thus constructed an RPF from its absoulte probability function without loss of information.

6 From Outcomes to Events

Our next task is to upgrade P to operate on the event level. This is more difficult than it seems. For example, we may wish to declare that the probability of event e_1 with respect to e_2 is going to be additive on e_1 as follows:

$$P(e_1, e_2) = \sum_{h_1 \in e_1} P(h_1, e_2) \tag{2}$$

Equation 2 looks uncontroversial, but it actually contradicts the fundamental axioms! If we let $e_1 = \varnothing$, then we have an empty sum on the right hand side of the equation, and we get $P(\varnothing, e_2) = 0$. Likewise, if we allow e_2 to be empty, we get $P(e_1, \varnothing) = P(\varnothing, e_1)^{-1} = 0^{-1} = \infty$. Both of these statements make sense until you realize that $P(\varnothing, \varnothing) = 0 = \infty$, and what's worse is that they are also equal 1 under the identity axiom!

Another problem arises when events are *internally non-comparable*, meaning that event e contains outcomes h_1 and h_2 where $P(h_1, h_2) = *$. Perhaps there are still a few interesting things we can say about such an event, but here we will constrain ourselves entirely to totally comparable outcome spaces in order to avoid such questions.

Definition 6.1. Let P be a totally comparable finite RPF. P can also measure the probability of two events relative to each other using the following rules:

(i) $P(e_1, e_2)$ obeys the fundamental axioms of relative probability.

(ii) $P(e_1, e_2)$ sums over any reference outcome h, so long as the result isn't indeterminate.

$$P(e_1, e_2) :\cong \frac{\sum_{h_1 \in e_1} P(h_1, h)}{\sum_{h_2 \in e_2} P(h_2, h)} \tag{3}$$

Because we no longer have access to absolute probability, the best we can do is measure it relative to a *reference outcome* h^*. This ratio might be indeterminate, so we use the matching relation instead of equality. Fortunately, we can show that there exists at least one reference outcome that will constrain $P(e_1, e_2$ in 3 if they are non-empty.

Proof. Each event in a totally comparable RPF must have an anchor internally, using the same argument made for proving the existance of anchors in lemma 5.3. Choose an internal anchor a from one of the events, say e_1. Then the sum $\sum_{h_1 \in e_1} P(h_1, a)$ will be non-infinite by definition of anchors, and non-zero because $P(a, a) = 1$ is a term in the sum. Therefore, the constraint as a whole cannot be indeterminate.

If both events are empty, then we are unable to create an anchor element, but by the identity axiom $P(\varnothing, \varnothing) = 1$. □

These requirements again seem reasonable, but how can we know for sure that they provide a complete and consistent definition of $P : \mathcal{F} \times \mathcal{F} \to \mathbb{M}^*$? The following must be shown:

(i) If two distinct values for h^* in statement 3 yield constraints on P, then they must be equal.

(ii) The constraint in 3 will not violate the fundamental axioms.

Proof. For (i):

Let r_1 and r_2 be distinct reference outcomes, and both constrain $P(e_1, e_2)$. Then we want to check that

$$\frac{\sum_{h_1 \in e_1} P(h_1, r_1)}{\sum_{h_2 \in e_2} P(h_2, r_1)} = \frac{\sum_{h_1 \in e_1} P(h_1, r_2)}{\sum_{h_2 \in e_2} P(h_2, r_2)} \tag{4}$$

Neither expression is a wildcard, and none of the individual terms are either. The key to this argument is in looking at the value of $P(r_1, r_2)$.

Assume $P(r_1, r_2) = 0$.

Then if $\sum_{h_1 \in e_1} P(h_1, r_1)$ is not infinite, then $\sum_{h_1 \in e_1} P(h_1, r_2)$ must be zero. The same argument applies to $\sum_{h_2 \in e_2} P(h_2, r_2)$. Since they can't both be zero, we can say that one of the sums on the left hand

side is infinite, so that $P(e_1, e_2)$ is either ∞ or 0. Let's say it is 0. Then $\sum_{h_1 \in e_1} P(h_1, r_1) = 0$ and $\sum_{h_2 \in e_2} P(h_2, r_1) = \infty$ and by the argument above $\sum_{h_1 \in e_1} P(h_1, r_2) = 0$. Because the right hand side is not $*$ - it must resolve to zero as well. The same agument holds for $P(e_1, e_2) = \infty$.

By an analogous argument, if $P(r_1, r_2)\infty$ then equation 4 must hold.

So now we can assume that $P(r_1, r_2) \notin 0, \infty$. The allows us to multiply the left hand side of equation 4 by $1 = \frac{P(r_1,r_2)}{P(r_1,r_2)}$ and distribute into the sum to get:

$$\frac{\sum_{h_1 \in e_1} P(h_1, r_1) \cdot P(r_1, r_2)}{\sum_{h_2 \in e_2} P(h_2, r_1) \cdot P(r_1, r_2)} = \frac{\sum_{h_1 \in e_1} P(h_1, h_2^*)}{\sum_{h_2 \in e_2} P(h_2, h_2^*)}$$

For (ii):

The identity, inverse, and composition axioms easily follow from the fact that equation 3 is a ratio with identical terms for e_1 in the numerator and e_2 in the denominator. Therefore, if it resolves it is just a ratio of positive numbers - which can be shown to follow the 3 axioms. □

Theorem 6.1. *Given events e_1 and e_2 where they are not both empty, we have*

$$P(e_1, e_2) = \sum_{h_1 \in e_1} \frac{1}{\sum_{h_2 \in e_2} P(h_2, h_1)}.$$

Proof. Start with the equation and multiply and use a suitable reference outcome h.

$$\sum_{h_1 \in e_1} \frac{1}{\sum_{h_2 \in e_2} p(h_2, h_1)} :\cong \sum_{h_1 \in e_1} \frac{P(h_1, h)}{\sum_{h_2 \in e_2} P(h_2, h_1) P(h_1, h)} = \frac{\sum_{h_1 \in e_1} P(h_1, h)}{\sum_{h_2 \in e_2} P(h_2, h)}$$

Since both $P(e_1, e_2)$ and the formula above match the same thing which is not always $*$, they must be equal. □

We then derive the absolute probability function as

$$P(e) = P(e, \Omega) = \sum_{h \in e} \frac{1}{\sum_{h' \in \Omega} p(h', h)}$$

Theorem 6.2. *The empty event $\varnothing$ has probability 0 with respect to any non-empty event.*

Proof. Let e be a non-empty event, and let h be an outcome in e.

$$P(\varnothing, e) :\cong \frac{\sum_{h_1 \in \varnothing} P(h_1, h)}{\sum_{h_2 \in e} P(h_2, h)} = \frac{0}{\sum_{h_2 \in e} P(h_2, h)}$$

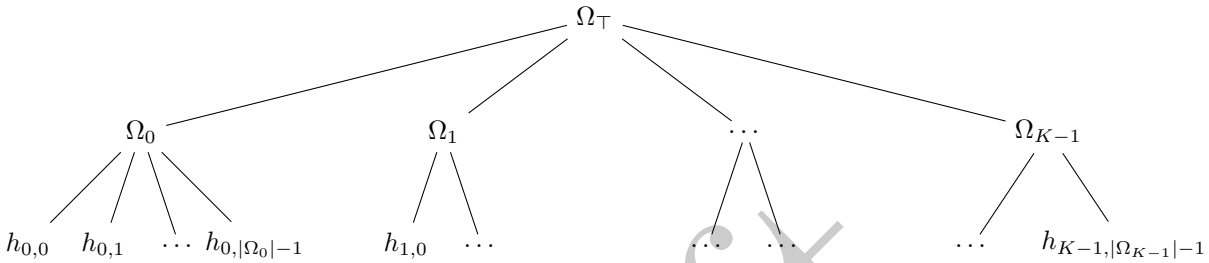
The denominator cannot itself be zero because $P(h, h)$ is one of the terms. Therefore, $P(\varnothing, e) = 0$ □

7 Composing Relative Probability Functions

Let $P_0, P_1, ..., P_{K-1}$ be relative probability functions. Each of these probability functions have unique categories in their own right.

Let the set of outcomes acted upon by P_k be Ω_k, so that $P_k : \Omega_k \times \Omega_k \to \mathbb{M}^*$. We take all Ω_k to be disjoint from one another.

We can combine all of these relative probability functions together with a top level probability function $P_\top$[10] with outcome space $\Omega_\top = \{\Omega_0, \Omega_1, \Omega_2, ...\Omega_{K-1}\}$.



Now let Ω be the set of all outcomes $\Omega_0 \cup \Omega_1 \cup \ldots \Omega_{K-1}$. We can create a new RPF - just called P acting on Ω - with the following assumptions:

1) If the two outcomes fall under the same component, then their relative probabilities do not change:

$$P(h_{k,i}, h_{k,j}) = P_k(h_{k,i}, h_{k,j}) \tag{5}$$

2) If the two outcomes fall under different components, then their relative probabilities are given as follows.

$$P(h_{k_1,i}, h_{k_2,j}) = P_{k_1}(h_{k_1,i}, \Omega_{k_1}) \cdot P_\top(\Omega_{k_1}, \Omega_{k_2}) \cdot P_{k_2}(\Omega_{k_2}, h_{k_2,j}) \tag{6}$$

Note the use of the composition property to traverse up and down the tree. One could of course imagine this for a tree being many levels, and having a different height for each branch.

Theorem 7.1. *P respects the fundamental axioms.*

Proof. Identity is obvious because an outcome is on the same component as itself, so we can use equation 5 to get $P(h_{k,i}, h_{k,i}) = P_k(h_{k,i}, h_{k,i}) = 1$

The inverse and composition laws must be true if both inputs are in the same component, because that component already follows the axioms. We not assume that the two inputs are from different components.

[10] Pronounced "P-Top".

The inverse law can be proven by calculation.

$$\begin{aligned} P(h_{k_1,i}, h_{k_2,j})^{-1} &= (P_{k_1}(h_{k_1,i}, \Omega_{k_1}) \cdot P_\top(\Omega_{k_1}, \Omega_{k_2}) \cdot P_{k_2}(\Omega_{k_2}, h_{k_2,j}))^{-1} \\ &= P_{k_1}(h_{k_1,i}, \Omega_{k_1})^{-1} \cdot P_\top(\Omega_{k_1}, \Omega_{k_2})^{-1} \cdot P_{k_2}(\Omega_{k_2}, h_{k_2,j})^{-1} \\ &= P_{k_1}(\Omega_{k_1}, h_{k_1,i}) \cdot P_\top(\Omega_{k_2}, \Omega_{k_1}) \cdot P_{k_2}(h_{k_2,j}, \Omega_{k_2}) \\ &= P_{k_2}(h_{k_2,j}, \Omega_{k_2}) \cdot P_\top(\Omega_{k_2}, \Omega_{k_1}) \cdot P_{k_1}(\Omega_{k_1}, h_{k_1,i}) \\ &= P(h_{k_2,j}, h_{k_1,i}) \end{aligned} \tag{7}$$

Composition can be shown similarly - now naming the 3 separate indecies in components k_1, k_2, k_3 as i_1, i_2, i_3 respectively.

$$\begin{aligned} &P(h_{k_1,i_1}, h_{k_2,i_2}) \cdot P(h_{k_2,i_2}, h_{k_3,i_3}) \\ &:\cong P_{k_1}(h_{k_1,i_1}, \Omega_{k_1}) \cdot P_\top(\Omega_{k_1}, \Omega_{k_2}) \cdot P_{k_2}(\Omega_{k_2}, h_{k_2,i_2}) \cdot P_{k_2}(h_{k_2,i_2}, \Omega_{k_2}) \cdot P_\top(\Omega_{k_2}, \Omega_{k_3}) \cdot P_{k_3}(\Omega_{k_3}, h_{k_3,i_3}) \\ &:\cong P_{k_1}(h_{k_1,i_1}, \Omega_{k_1}) \cdot P_\top(\Omega_{k_1}, \Omega_{k_2}) \cdot P_\top(\Omega_{k_2}, \Omega_{k_3}) \cdot P_{k_3}(\Omega_{k_3}, h_{k_3,i_3}) \\ &:\cong P_{k_1}(h_{k_1,i_1}, \Omega_{k_1}) \cdot P_\top(\Omega_{k_1}, \Omega_{k_3}) \cdot P_{k_3}(\Omega_{k_3}, h_{k_3,i_3}) \\ &:\cong P_{k_1}(h_{k_1,i_1}, h_{k_3,i_3}) \end{aligned} \tag{8}$$

□

Theorem 7.2. *P is totally comparable if and only if the following are true:*

1. *$P_\top$ is totally comparable.*
2. *For all $k \in \{0, 1, ..., K-1\}$, P_k is totally comparable.*
3. *All components except at most one are totally mutually possible.*
4. *If there is a component that is not totally mutually possible, then every element of $P_\top$ possible with respect to that component.*

Proof. If all the components are totally comparable, then we only need to prove that outcomes in different components are comparable. Starting with equation 6,

$$P(h_{k_1,i}, h_{k_2,j}) = P_{k_1}(h_{k_1,i}, \Omega_{k_1}) \cdot P_\top(\Omega_{k_1}, \Omega_{k_2}) \cdot P_{k_2}(\Omega_{k_2}, h_{k_2,j}) \tag{9}$$

The only way that we can get $P(h_{k_1,i}, h_{k_2,j}) = *$ is if there are both 0 and ∞ as factors on the right hand side.

Because there is at most one component with outcomes impossible with respect to that component, we can say that either $P_{k_1}(h_{k_1,i}, \Omega_{k_1}) = 0$ or $P_{k_2}(h_{k_2,j}, \Omega_{k_2}) = 0$, or possibly neither, but not both.

Neither can be infinite either by the definition of the event level in equation 3. Here we look at the factor $P_{k_1}(h_{k_1,i}$ and use k_1 itself as the reference outcome.

$$P_{k_1}(h_{k_1,i}, \Omega_{k_1}) :\cong \frac{\sum_{h_1 \in \{k_1\}} P(h_1, k_1)}{\sum_{h \in \Omega_{k_1}} P(h_2, k_1)} = \frac{1}{\sum_{h \in \Omega_{k_1}} P(h_2, k_1)}$$

The denominator cannot be zero since $P(k_1, k_1) = 1$ will be one of the terms under the sum.

If the $P_{k_1}(h_{k_1,i} = 0$, then the only way the entire right hand side can be $*$ is if $P_\top(\Omega_{k_1}, \Omega_{k_2}) = \infty$. But this can't be true because we assumed that Ω_{k_2} is possible with respect to Ω_{k_1}, the sole component with impossible outcomes!

An analogous argument can be made if $P_{k_2}(h_{k_2,j}, \Omega_{k_2} = 0$.

Therefore, the right hand side of the equation is not $*$ and P is totally comparable.

In the opposite direction, we can show that if any of the conditions are broken, then P is not totally comparable. Breaking any of the first two conditions would introduce an explicit $*$ into equation 6. If there are multiple components with impossible outcomes, then it would introduce a 0 into the first term of equation 6 and an ∞ into the third term, yielding $*$.

And finally, if only the fourth condition is broken, it would introduct a 0 into the first term of equation 6 and an ∞ into the **second** term of equation 6.

Therefore, if any of these conditions are broken, P is **not** totally comparable. □

8 Bayesian Inference on Relative Distributions

A relative probability function represents a belief over the set of potential hypotheses in Ω.

Start with the Bayesian inference formula for conditional probability for $h \in \Omega$ assuming that we recieve data D.

$$P(h|D) = \frac{P(D|h) \cdot P(h)}{P(D)} \qquad P(D) = \sum_{h \in \Omega} P(D|h) \cdot P(h)$$

Now we convert to relative probability by looking at the ratio between the two hypotheses.

$$\frac{P(h_1|D)}{P(h_2|D)} = \frac{P(D|h_1) \cdot P(h_1)}{P(D)} \div \frac{P(D|h_2) \cdot P(h_2)}{P(D)} = \frac{P(D|h_1) \cdot P(h_1)}{P(D|h_2) \cdot P(h_2)}$$

Notice that each component is represented by a ratio. By making the appropriate subsitutions, we can express this entirely in terms of relative probability functions.

For the ratio of prior probabilities, substitute the relative prior: $\frac{P(h_1)}{P(h_2)} \to P(h_1, h_2)$

For the ratio of posterior probabilities, substitute the relative posterior: $\frac{P(h_1|D)}{P(h_2|D)} \to P(h_1, h_2|D)$

It is more difficult to see that the likelihood ratio is a relative probability, but the Kolmogorov definition to expand conditional probability suggests that it is:

$$\frac{P(D|h_1)}{P(D|h_2)} = \frac{\frac{P(D \cap h_1)}{P(D)}}{\frac{P(D \cap h_2)}{P(D)}} = \frac{P(D \cap h_1)}{P(D \cap h_2)}$$

Therefore, we can let P_D represent the likelihood ratio of the different hypotheses, and we can be sure that it fits the RPF framework with regards to the fundamental axioms. The likelihood ratio $P_D(h_1, h_2)$ encodes a description of how the different hypotheses rate the likelihood of data.

The substitution for the liklihood ratio is as follows: $\frac{P(D|h_1)}{P(D|h_2)} \to P_D(h_1, h_2)$

Now we get bayes rule for relative probability:

$$P(h_1, h_2|D) = P_D(h_1, h_2)P(h_1, h_2)$$

Bayesian inference is not reduced to an element-by-element multiplication of two different RPFs: $P_D(h_1, h_2)$ and $P(h_1, h_2)$. Fortunately, product of two RPFs also obeys the fundamental axioms.

Theorem 8.1. *Let P_1 and P_2 be relative probability functions on Ω. Define $P(h_1, h_2) = P_1(h_1, h_2) \cdot P_2(h_1, h_2)$. Then, P is also an RPF, that it is obeys the fundamental axioms.*

Proof. Use the multiplication property of the matching relation in equation 2.5.

Identity:

$$P(h_1, h_1) = P_1(h_1, h_1)P_2(h_1, h_1) = 1 \cdot 1 = 1$$

Inverse:

$$P(h_1, h_2) = P_1(h_1, h_2) \cdot P_2(h_1, h_2) = P_1(h_2, h_1)^{-1} \cdot P_2(h_2, h_1)^{-1} = (P_1(h_2, h_1) \cdot P_2(h_2, h_1))^{-1} = P(h_2, h_1)^{-1}$$

Composition:

$$P(h_1, h_2)P(h_2, h_3) = P_1(h_1, h_2)P_2(h_1, h_2)P_1(h_2, h_3)P_2(h_2, h_3) :\cong P_1(h_1, h_3)P_2(h_1, h_3) =$$

□

Theorem 8.2. *Once two outcomes become uncomparable, they will never be comparable again. In other words, if $P(h_1, h_2) = *$, then $P(h_1, h_2|D) = *$.*

Proof. $P(h_1, h_2|D) = L(D|h_1, h_2)P(h_1, h_2) = P_D(h_1, h_2) \cdot * = *$ □

Theorem 8.3. *Once an outcome becomes impossible with respect to another event, it will either remain impossible or become uncomparable. In other words, if $P(h_1, h_2) = 0$, then $P(h_1, h_2|D) \in 0, *$.*

Proof. $P(h_1, h_2|D) = L(D|h_1, h_2)P(h_1, h_2) = P_D(h_1, h_2) \cdot 0$. Normally, this would simplify to 0, but with the matching relation in $\mathbb{M}^*$, this will be $*$ if $P_D(h_1, h_2) \in \infty, *$. □

8.1 Example: A Noisy Channel

Here is an example of how relative probability gives us an interesting way of looking at inference problems.

Suppose we are to recieve a message in outcome space $\Omega = \{0, 1, ..., K-1\}$. There is a probability of p that the message goes through correctly. Otherwise, it gets scrambled and we recieve a value in Ω drawn from the uniform distribution[11]. We recieve the same message several times for redundancy, and count c_i as the number of times the message was recieved as k.

To start with absolute probability, we can use the indicator function to get the probability of recieving h_1 given that the real message was h_2 is $p[h_1 = h_2] + \frac{1-p}{K}$. We then use this to construct an RPF for the liklihood ratio if we recieve a single message, k.

$$P_k(h_1, h_2) = \frac{p[h_1 = k] + \frac{1-p}{K}}{p[h_2 = k] + \frac{1-p}{K}} = \frac{pK[h_1 = k] + 1 - p}{pK[h_2 = k] + 1 - p}$$

Now if we recieve multiple messages in the count vector c:

$$P_c(h_1, h_2) = \prod_{k \in \Omega} \left(\frac{pK[h_1 = k] + 1 - p}{pK[h_2 = k] + 1 - p} \right)^{c_k}$$

Note that if $k \notin \{h_1, h_2\}$ then the term for that k becomes $\frac{1-p}{1-p} = 1$, so there are only terms that we care about. We will also assume $h_1 \neq h_2$:

$$\begin{aligned} P_c(h_1, h_2) &= \left(\frac{pK[h_1 = h_1] + 1 - p}{pK[h_2 = h_1] + 1 - p} \right)^{c_{h_1}} \left(\frac{pK[h_1 = h_2] + 1 - p}{pK[h_2 = h_2] + 1 - p} \right)^{c_{h_2}} \\ &= \left(\frac{pK + 1 - p}{1 - p} \right)^{c_{h_1}} \left(\frac{1 - p}{pK + 1 - p} \right)^{c_{h_2}} = \left(1 + \frac{pK}{1 - p} \right)^{c_{h_1} - c_{h_2}} \end{aligned} \tag{10}$$

Because the prior is uniform, we also get the posterior:

$$P(h_1, h_2 | c) = P_c(h_1, h_2) \cdot P(h_1, h_2) = \left(1 + \frac{pK}{1 - p} \right)^{c_{h_1} - c_{h_2}}$$

Note that the relative probability between two hypotheses is exponential on the difference between their counts. Formulating these problems in terms of relative probability often lead to easily interpretable results, even before converting into absolute probability (if that is even required). Using a different prior would be as easy as appending an additional term.

9 Implementation

Finally, we implement relative probabiliy as a python class as a demonstration of its usage and relevance.

How to implement this in code, and point to open source example.

Note the connection between magnitude space and the extended real number line, which we can implement through floating point numbers.

This can be implemented by storing K values.

[11] We could still have gotten lucky and recived the correct value in this case

For each category, we have a tier. Items in the same tier are comparable. Each Tier has a parent tier, where items in this teir are said to be impossible relative to anything in its ancestor tiers.

For each category, we also store a floating point number called the value, which should be taken as the log of an unnormalized probability. Note that we will not allow inf or NaN here.

Get the relative probability of 2 categories. Algorithm: If they are in the same tier, then subtract their values and take the exp. If they are in different tiers, do a graph search on the tier. If the first is ¡ the second, the answer is 0. If the first is ¿ the second, the answer is 1. And if they are uncomparable, then the answer is Wildcard, NaN.

Generate and indifferent distribution of category K. Algorithm: Create a single tier where all values are set to 0.

Change the relative probability of item k_1 with respect to k_2, and set it to q. Algorithm: UNSURE

Set the probability k_1 to some absolute value with respect to either the whole distribution, or to its tier.

Randomly sample from this distribution. Algorithm: Only look at the top tier.

Randomly sample from this distribution, but remove certain categories. Algorithm: If the top tier categories are gone, look to see if a top tier remains. If there are multiple top tiers, then there's no way to do it!

Ask: Is this distribution totally mutually possible? Algorithm: Look are a single top tier.

Ask: Is it totally comparable? Algorithm: Look for a linear list of tiers.

10 Topology and Limits in Relative Probability Space

Mathematics can model the real world even through ideas that are theoretically impossible. For example, we might believe that a certain natural process cannot repeat an infinite number of times - that it just is not something allowed by the physical limitations of our universe. And even so, we might still take its value to be infinity in a model in order to get a bound on what that system will look like in "the long run". One of the benefits of relative probability spaces is their properties with respect to limits. To this end, we prove here that when we take limits of totally comparable RPFs, the resulting RPF will also be totally comparable.

This effort caps off the most significant argument in favor of totally comparable RPFs. They hold their information under the operation of limits, while absolute probability does not.

There is some background in topology[12] required for this section.

10.1 RPF Space and Compactness

Because the set of absolute distributions for K is embedded in $\mathbb{R}^K$, its topological properties are well understood. The simplex is closed, bounded, and compact. Practically, this means that any sequence of points on

[12]See Mendelson (1990) [3] and Bradley et al. (2020) [4] for texts with formal definitions and theorems.

the simplex will converge to one or more points on the simplex allowing both pure and applied practicioners to talk about limit and boundary conditions.

This strategy fails for relative probabilities, because there is no obvious way to embed an RPF into euclidean space[13]. The relative probability space is more complicated, because at the corners of the simplex lurk entire subspaces where outcomes are still being compared in different ways!

Definition 10.1. $\text{RPF}^*(K)$ is the set of relative probability functions of size K (where $\Omega = \{0, 1, ..., K-1\}$). Likewise $\text{RPF}(K)$ is the set of all totally comparable RPFs of size K.

If the $\text{RPF}(K)$ is compact, then information about the relative probabilites of events are preserved even as they approach zero relative to another event.

In order to prove compactness, we first must define a topology on $\text{RPF}(K)$. This starts with finding a *basis of open sets*.

The notion of an open set can change even if a topological space is restricted. For example, on the real number line $\mathbb{R}$, we take the open interval (0, 1) as an open set (as the term open interval suggests). However, once this is embedded into $\mathbb{R}^2$, it is now a line segment in a plane and no longer open (see figure 6). It can be thought of as a restriction to an open set on $\mathbb{R}^2$ to $\mathbb{R}$. For example, the set $\{(x, y) : x \in (0, 1) \text{ and } y \in (-\epsilon, +\epsilon)\}$ given an $\epsilon > 0$ is such an open set on $\mathbb{R}^2$.

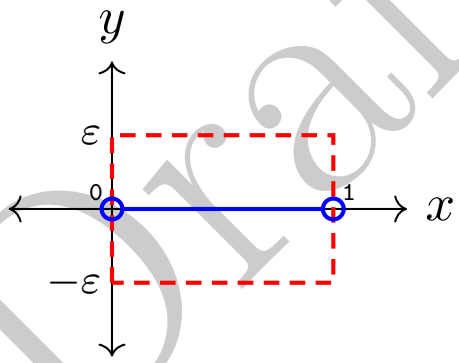


Figure 6: The small box that is the interior of the dotted rectangle is an open set in $\mathbb{R}^2$, and therefor its restriction to $\mathbb{R}$ - the line segment - is an open set in $\mathbb{R}$. But the line segment is not open in $\mathbb{R}^2$.

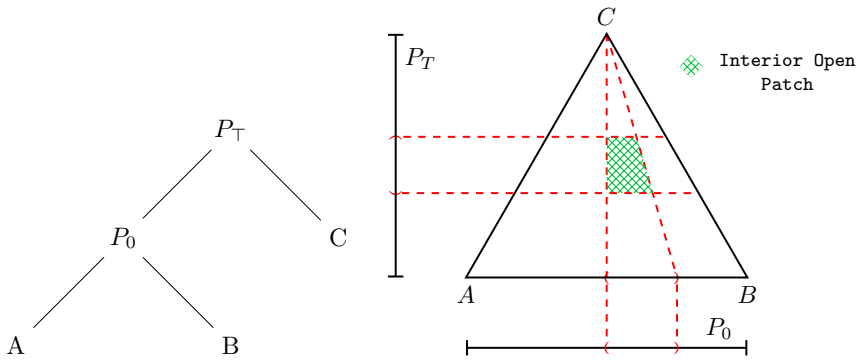
Likewise, an open set on a relative probability space restricted on several outcomes might not be an open set on the relative probability spaces for all of Ω.

We start by looking at RPFs with $K = 2$. Fortunately, we find a totally comparable RPF that corresponds 1:1 with the magnitude space.

Theorem 10.1. *Let $\Omega = \{h_1, h_2\}$ have two elements, with relative probability function P. Then, P is completely determined by $P(h_1, h_2)$.*

Proof. Let $q = P(h_1, h_2)$. By the inverse symmetric property, $P(h_2, h_1) = q^{-1}$. These values completely determine P on the outcome level. □

[13]Though it may be possible! See section 11.5



This gives us both a topology and a compactness proof for K = 2 for free because RPF(2) is isomorphic to $\mathbb{M}$ which already has a natural topology. Its basis for open sets are the open intervals of $\mathbb{B}$, including those intervals that include 0 and ∞. For $K > 2$, we will need more powerful tools.

10.2 Open Patches

We now develop a notion of open patches, which will be a basis of open sets on the space RPF(K).

Definition 10.2. An *interior open patch* of $\text{RPF}_{\text{comp}}(\Omega)$ is one of the following:

1. If $K = 2$, a subset parameterized by an interior open interval of magnitudes. $\{P|a < P(h_1, h_2) < b\}$ for some $a, b \in \mathbb{M}$
2. If $K > 2$, a composition of interor patches with composing function $P_\top$ also being an interior patch.

Intuition: Interior open patches contain only totally mutually possible functions.

Definition 10.3. A *facet*[14] *patch* of $\text{RPF}_{\text{comp}}(\Omega)$ is one of the following:

1. If $K = 2$, an interval of the form $\{P|0 < P(h_1, h_2) < a\}$ for some $a \in \mathbb{M}$
2. If $K > 2$, the set of compositions where $P_\top$ is drawn from an interior open patch, and all but one of the components are drawn from interior open patches. The last component - called the *facet component* is itself a facet patch.

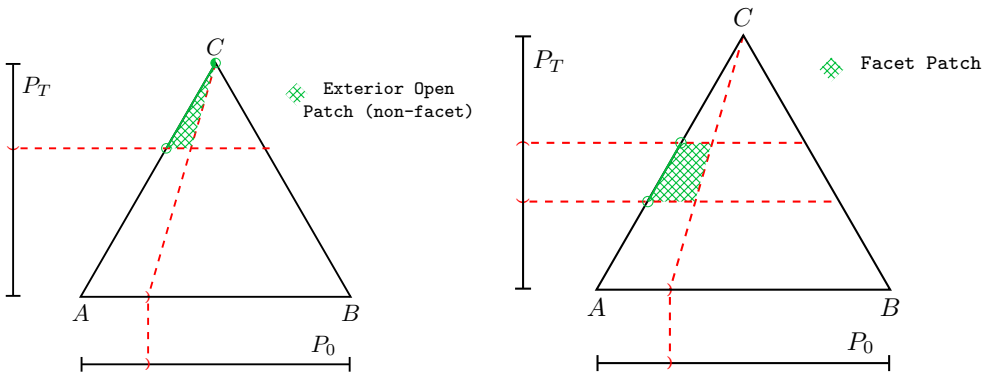
Definition: An *exterior open patch* is a one of the following:

1. Any facet patch.
2. A composition where $P_\top$ is a facet patch. The *facet component* is itself drawn from any open patch, and all the other components are drawn from interior open patches.

Exterior open patches contain only totally mutually comparable functions, but some are not totally mutually possible.

TODO: Break this down because it's not that intuitive!

[14] A facet of a simplex is a subset where one parameter is equal to zero - equivalent to a face on a 3D object.



Definition 10.4. An *open patch* is a subset of RPF(K) that is either an interior or exterior open patch.

Now let the open patches be the bases for an open set thus defining a topology on RPF(K).

10.3 Compactness

Before we prove that RPF(K) is compact, we need a few lemmas.

Lemma 10.2. *Let h be an outcome in $\{0, 1, ..., K-1\}$, and let q be a number such that $0 < q \leq 1$. The region of RPF(K) where the absolute probability of h is q is isomorphic to RPF($K-1$).*

Proof. Roughly: If $P(h) > 0$ then it is in the anchored equivalence class of mutually possibility. If $P(h) = 1$ then the outcome h can be appended above any function in RPF($K-1$), and if $P(h) < 1$ then h can be appended into the anchored equivalence class of any function in RPF($K-1$). In both cases, a separate h with a given absolute probability can be appended to anything in RPF($K-1$) to produce an element of RPF(K), and all elements of RPF(K) are accounted for. □

Lemma 10.3. *Let h be an outcome in $\{0, 1, ..., K-1\}$. Let R be the region of RPF(K) where the absolute probability of h is 0. Any open set containing all of R must contain every function in RPF(K) where $P(h) < \epsilon$ for some $\epsilon < 0$.*

Theorem 10.4. *RPF(K) is compact, meaning that for every open cover of it, there is a finite subcover.*

Proof. This is an inductive proof where we assume that the theorem is true for all $k < K$ and then prove that it is true for K.

If $K \in 0, 1$ then RPF(K) is finite and singular (either the empty RPF or unit RPF respectively). These are obviously compact. If $K = 2$ then we have the topology of $\mathbb{M}$ which is also compact (thanks to the ∞ element).

Now we assume that $K > 2$.

Let $h \in \Omega$ be an outcome.

- We're going to have to prove this - might be tough! □

10.4 Simple Limit Example

Let us define a simple relative probability distribution P_q where $K = 3$ that is parameterized by the magnitude $q \in \mathbb{M}$.

Let $P_q(h_0, h_1) = q$ and $P_q(h_1, h_2) = 2$.

By the fundamental property, $P_q(h_0, h_2) :\cong P_q(h_0, h_1) \cdot P_q(h_1, h_2) = 2q$.

Now we want to consider the case where the relative probability of h_0 grows infinitely large in comparison to h_1 and h_2.

$$P = lim_{q \to \infty} P_q$$

We use the following topological definition for the limit in this case: For every open set A of relative probability distributions containing P, there exists an open interval $B = (b, \infty)$ on $\mathbb{M}$ such that for every value of $q \in B$, P_q is in A.

Proposition 10.5. *The above limit that defines P exists, and $P(h_1, h_2) = 2$. In other words, h_2 is still half as likely as h_1 and that information hasn't been lost on P.*

Proof. TODO □

11 Future Work

11.1 Expansions to infinite spaces

- Including topological and metric - Much richer world, more complex mathematics, more applications - Is it possible to create a univified version of the Hausdorff measure, where objects are categorized by dimention d, and a smaller-dimentional object is always mutually impossible to a larger dimentional object.

11.2 Connection Surreal Numbers

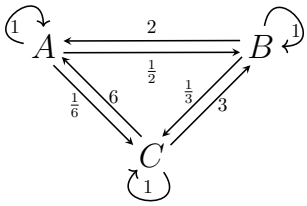
- This is greater, richer than the real number system - Does this abrograte the need for the relative probability function (not for incomparable values) - If the infinite case is dealt with above, then more questions are raised about both the power of surreal numbers and their suitability

11.3 Shrinking the Measure Number System

- We still have a usable system if we want Rational Numbers - Can this system work for all non-standard probability value systems? - There is practical application in this work, since computers cannot work with real numbers directly. We implement this system with floating point numbers and this approximation should be good enough for most applications - but can we have a version with more precise arithmetic

11.4 Relationship to Category Theory

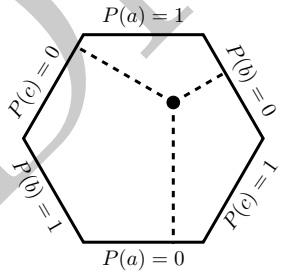
Category theorists will instantly recognize that an RPF describes a category perfectly. This construction can be analyzed and approached through the lens of category theory.



The recent work of Censi et al.[7] concerns negative information in categories, which here corresponds to the wildcard element $*$. It represents regions of the probability function that remain unassigned or uncomparable. This work could be used to subsume and further develop the idea of the indeterminate wildcard.

11.5 Embedding in Euclidean Space

Absolute probability functions have this advantage where they can be embedded into a simplex in $\mathbb{R}^K$. For relative probability functions, it is not so straightforward. However, it might be possible to derive a system for embedding RPFs into euclidean space. For example, the space RPF(3) can be mapped as a hexagon, where each point can be assigned a probability based on its distance between two parallel sides, which exist for each outcome.



In this case, the probability triangle has been truncated. For higher order simplcies, this appears to become exceedingly unweidly unless there is developed some simplifying trick. If it is successfully done, then the topological properties of RPF(K) fall into place easily.

References

[1] Sklar, M. (2014). Fast MLE computation for the Dirichlet multinomial. arXiv preprint arXiv:1405.0099.

[2] Sklar, M. (2022). Sampling Bias Correction for Supervised Machine Learning: A Bayesian Inference Approach with Practical Applications. arXiv preprint arXiv:2203.06239.

[3] Mendelson, B. (1990). Introduction to topology. Courier Corporation.

[4] Bradley, T. D., Bryson, T., & Terilla, J. (2020). Topology: A Categorical Approach. MIT Press.

[5] Lyon, A. (2016). Kolmogorov's Axiomatisation and its Discontents. The Oxford handbook of probability and philosophy, 155-166.

[6] Hájek, A. (2003). What conditional probability could not be. Synthese, 137(3), 273-323.

[7] Censi, A., Frazzoli, E., Lorand, J., & Zardini, G. (2022). Categorification of Negative Information using Enrichment. arXiv preprint arXiv:2207.13589.

[8] Kahan, W. (1996). IEEE standard 754 for binary floating-point arithmetic. Lecture Notes on the Status of IEEE, 754(94720-1776), 11.

[9] A. N. Kolmogorov. Foundations of the Theory of Probability. Chelsea Publishing Company, New York (1956).

[10] Heinemann, F. (1997). Relative Probabilities. Working paper, http://www. sfm. vwl. uni-muenchen. de/heinemann/publics/relative probabilities-intro. htm.

[11] Matoušek, J., & Nešetřil, J. (2008). Invitation to discrete mathematics. OUP Oxford.

This document along with revisions is posted at github as https://github.com/maxsklar/relative-probability-finite-paper. See readme for contact information. Local Maximum Labs is an ongoing effort create an disseminate knowledge on intelligent computing.