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 explore!

Since Kolmogorov published the standard axioms for probability[9] in 1933, there have been calls to alter them for various applications. In Kolmogorov's Axiomatisation and Its Discontents[5], Lyon lays out these cases and their justifications. One area of "discontentment" concerns conditional probability. We often want to identify the probability of event A given event B - or $P(A|B)$ - but can this be done if B has probability zero[1]?

We are out of luck with the Kolmogorov model, which defines $P(A|B)$ as the ratio $\frac{P(A \cap B)}{P(B)}$. When $P(B) = 0$, the indeterminate form $\frac{0}{0}$ appears leaving the conditional probability undefined. This will happen whenever one wishes to compare two events that each have an overall probability of zero.

Undeterred, mathematicians and engineers refer to relative probabilities of this type all the time. For example, if we consider a probability distributrion over $[0, 1]$ given by $2x$, we know that the value at $x = \frac{1}{2}$ is twice as much as the value 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-dimensional object - has probability 0 compared to the rest of the globe.

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

This improves on the Kolmogorov model - that starts with an absolute probability function - by solving both the conditional probability question and giving rise to new objects to study. Furthermore, the model fits nicely into many modern frameworks such as category theory and bayesian distribution sampling.

1.1 Previous Work and Goals

Leading probability theorists in the twentieth century have developed axiomatic systems for conditional probability, notably Renyi[10]. Kohlberg and Reny[15] introduced the idea of relative probability and applied it to game theory. Heinemann[11] further developed the idea of *relative probability measures* along with their axioms and definitions, and found applications in bayesian inference and economics[12].

This work will expand on the idea of relative probability in several ways.

First, as a proof of 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.

[1] Another unintuitive feature of probability theory is that zero probability events do indeed occur, particularly when given a continuous distribution. November**??** gave a more recent philosophical treatment of this phenomenon.

[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

Relative probability requires a new set of fundamental definitions, which we will construct without the distractions of infinite and continuous outcome spaces.

Our focus on probability at the outcome level will lead us to separate out three funamental axioms of relative probability from those related to summation and measurability. These fundamental axioms establish relative probability as a thin category on an outcome space, ultimately provide a bridge to category theory.

Second, we will categorize the various patterns and states that arise when dealing with relative probability, including total comparability, possibility classes, and anchor outcomes.

Third, we focus on the computational properties of bayesian inference on relative probabilities. The relative probability function will simplify the formulas for specific distributions in the bayesian framwork. We introduce the notion of the indeterminate wildcard value, which will inevitably arise when relative probability is inferred on messy, real world data. We discuss ways in which the relative probability concept can be applied to code and data structures.

Finally, we discuss the ability of relative probability functions to retain information when taking limits.

Ultimately, practicioners will find these features of relative probability attractive. Its sphere of application could easily be expanded beyond game theory and theoretical probability into fields like machine learning and category theory. This paper aspires to provide foundational analysis that future researchers need to expand that sphere.

2 Preliminaries

2.1 Magnitude Space

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

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

Magnitudes correspond with our intuition of size. The value of infinity is a *limit element*, larger that all of the other magnitudes. It endows the magnitude space with several important properties:

1. Compactness: Sequences that go off to infinity still a limit (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 say there are infinite odds when an event with probability > 0 is compared to an event with probability 0. It is also common to find the odds of an event and its converse as its "odds." In this case, ∞ corresponds to events that are certain.
3. The infinite element is introduced in measure theory because many mathematical systems (real and natural numbers for example) contain sets of infinite measure.

We set $0^{-1} = \infty$ and $\infty^{-1} = 0$, even though the product $0 \cdot \infty$ is indeterminate.

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.

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 unique discipline:

- 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.
- 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.

$$0 \cdot \infty = *$$

$$* + m = *$$

$$* \cdot m = *$$

There is a cost to the wildcard introduction in that we now lose some basic properties of sums and products. For instance, 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 adjusted 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.* The wildcard element represents every single value, but it cannot be represented by any specific value. It also represents a loss of information about the parameter.

We will need a few lemmas which 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.

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

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

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 $m_3 = *$. In other words, if any of the three values are *ast*, then $m_3 = *$. By property 2.2, the theorem holds.

For non-symmetric, we present a counterexample: $1 :\cong *$ 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.

Finally, we establish that the matching relation preserves most operations such as addition and multiplication.

Lemma 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 $|\Omega| = K$. 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)-dimensional object called a 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).

3.1 Events

An *event* is a set of outcomes, and by convention $\mathcal{F}$ is the set of all possible events. In general, $\mathcal{F}$ is not the entire power set of Ω, but when Ω is finite we can consider any subset $e \subseteq \Omega$ to be an event[7] without concern.

In the previous section, the probability function was defined on individual outcomes. We now define the probability function on an event. The probability of an event is the probability that any one of its outcomes

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

[7] We need not concern ourselves with defining a σ-algebra of measurable sets here.

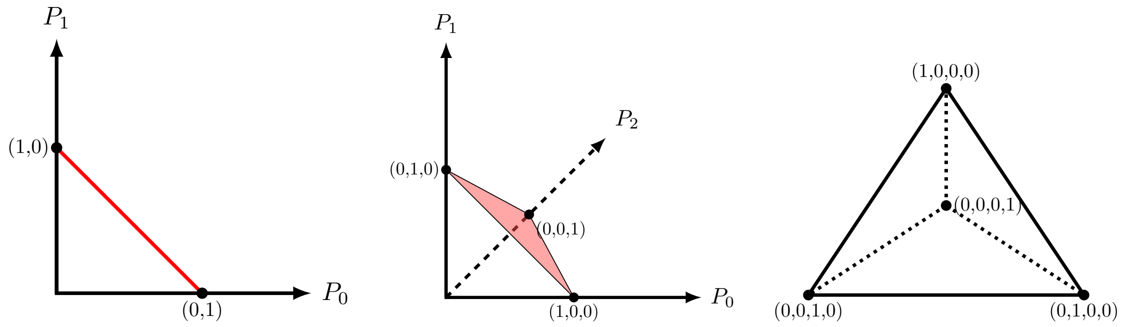


Figure 1: An illustration of the probability simplex for K = 2, 3, and 4. These objects are respectively, a segment embedded in $\mathbb{R}^2$, an equilateral triangle embedded in $\mathbb{R}^3$, and a normal tetrahedron embedded in $\mathbb{R}^4$. We make no attempt to visualize the 4D space that contains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

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 $\mathcal{F}$,

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

P acts on either outcomes or events using the obvious 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.

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.

With absolute probability, information is lost at the vertices where the probability of several outcomes might be assigned a value of zero. For example, if $\Omega = a, b, c$ with $P(a) = 1$ and $P(b) = P(c) = 0$, we cannot compar the probabilities of b and c by ratio as we can in the rest of the simplex.

This poses an interesting problem for limits.

Example 3.1. Consider the following categorical distribution function, with parameter $\epsilon > 0$:

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

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!

One of the most important properties of relative probabilities is their ability to compose as follows:

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 in Ω.

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*: $P(h_1, h_3) :\cong P(h_1, h_2) \cdot P(h_2, h_3)$

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

Let us pause for a moment to discuss how these axioms were 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 not, these relative probability assignments would have no meaning; they would just be numerical assignments without rhyme or reason[8].

The composition axiom is enough to show that the identity axiom works most of the time. For example, if h_1 is comparable 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 $P(h, h) = *$. The self-comparisons in an outcome space should not be able to contain any information where there is a choice of values. 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 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 constraint 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 $*$. 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 same piece of information but in reverse.

4.2 Examples

Now that the definition of relative probability is squared away, we can construct a library of examples for common RPFs that will serve as building blocks to tackling common problems.

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. It is as if the subjective probability agent gave up or the Bayesian model was fed corrupt data.

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 not the case for 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 $P(h, h) = 1$.

[8] Many of our political and economic forecasts come in this form.

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 n, and a probability of success p. The RPF has outcome space $\Omega = \{0, 1, 2, ..., n\}$ 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 Concepts for Relative Probability Functions

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

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

5.1 Absolute Probability Functions as a Special Case

Fortunately, the absolute probability function is a special case an RPF through definition 3.2, defined by $P(h_1, h_2) = \frac{P(h_1)}{P(h_2)}$ with the exception that if $P(h) = 0$, then $P(h, h) = 1$ instead of $*$ in order to satisfy the identity axiom.

This formula was shown to follow the composition axiom in theorem 3.1, and the inverse axiom follows easily from the formula as well.

5.2 Comparability, Possibility, and Anchors

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

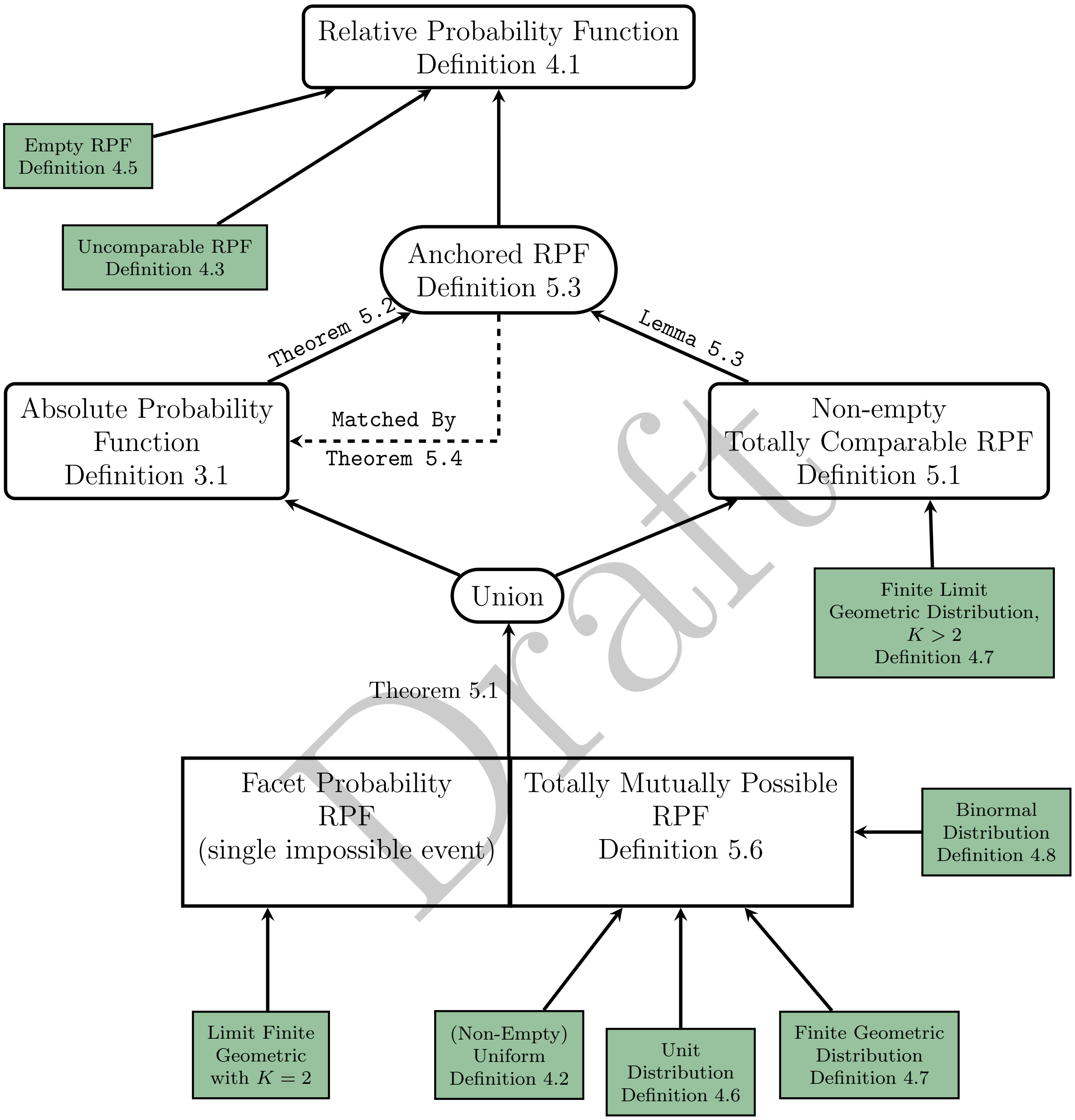


Figure 2: This is our roadmap for all of the sub-types of relative probability functions and their relationship to one another.

$P(h_2, h_1) = \infty$. Any other pairing that does not involve h_1 will be the quotient of two positive numbers, and thus is also comparable. □

With relative probability, it is important to know not only which outcomes are comparable, but whether their relative probability is positive and finite (that is, not equalt to 0, ∞, or $*$).

Definition 5.2. 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 h_2 if it is comparable and $P(h_1, h_2) > 0$.

Definition 5.3. An *anchor* of an RPF is an outcome that is possible with respect to every other outcome. An RPF that has at least 1 anchor is called an *anchored RPF*.

Anchor outcomes are those outcomes that have a non-zero absolute probability. The anchoring of a distribution ensures 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', $P(h, h') = \frac{P(h)}{P(h')} > 0$ □

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 for every outcome h, there exists another outcome h' such that $P(h, h') = 0$.

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

Let f^n be the function f applied 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 $i < N$.

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 its outcomes. Some RPFs have less information but are nevertheless consistent with RPFs that have more. The following definition encapulates this relationship.

Definition 5.4. 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. For all outcomes h_1 and h_2,

$$P_1(h_1, h_2) :\cong P_2(h_1, h_2)$$

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, $P(h', a) < \infty$. Therefore the sum $\sum_{h' \in \Omega} P(h', a) < \infty$. $\sum_{h' \in \Omega} P(h', a)$ is also > 0, because it includes the term $P(a, a) = 1$. The numerator $P(h, a)$ is also $< \infty$. Therefore, $P(h) \notin \{\infty, *\}$.

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 shown previously that they cannot be 0 or ∞.

Therefore, $P(h)$ is a valid absolute probability function. We show that the RPF is matched by it though the following calculation:

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

□

5.3 Mutual Possibility

Definition 5.5. Outcomes h_1 and h_2 are *mutually possible* if they are comparable and $0 < P(h_1, h_2) < \infty$. In other words, h_1 and h_2 are each possible with respect to the other.

Theorem 5.5. *Mutually possibility is an equivalence relation, being reflexive, symmetric and transitive.*

Proof. Reflexive: $P(h_1, h_1) = 1$ by the identity axiom.

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 is as well.

Transitive: 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 and finite, then their product is also positive and finite. □

Definition 5.6. An RPF is *totally mutually possible* if all of its outcomes[9] are mutually possible - and therefore all anchors.

It is helpful to make diagrams of possibility and impossibility through a *directed graph*. In these graphs, each outcome is represented by a point, and an arrow from A to B means that B is possible with respect to A. A bidirectional arrow means that A and B are mutually possible. Totally mutually possible RPFs have a simple diagram where all the outcomes are connected as in figure 3.

Theorem 5.6. *A non-empty totally mutually possible RPF is equal to an absolute probability function.*

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

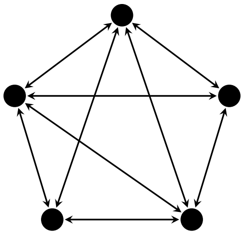


Figure 3: A totally mutually possible RFP has - unsurprisingly - a complete graph of mutually possibility.

Proof. If P is totally mutually possible, then all of its 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. □

5.4 Possibility Classes

In order to analyze the general case of RPFs, we need to consider classes of mutual possibility.

Theorem 5.7. *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*. Figure 4 is an example of outcomes grouped by mutually possible equivalence classes, with each class being impossible with respect to the ones it points to. Figure 4 is anchored while figure 5 is not anchored.

Finally, we look at totally comparable RPFs, where the graph of mutually possible components is a straight line (see figure 6).

Theorem 5.8. *If an RPF is totally comparable, then the equivalance classes of mutually possible outcomes are totally ordered. That is, the relative probability of outcomes from one class with respect to outcomes from another are always the same, and are either 0 or ∞.*

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.

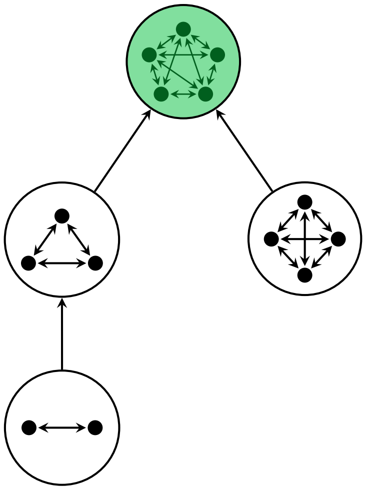


Figure 4: A diagram of an anchored RPF with its mutually possible classes. The anchor class (shaded) is the maximal class in the partial order.

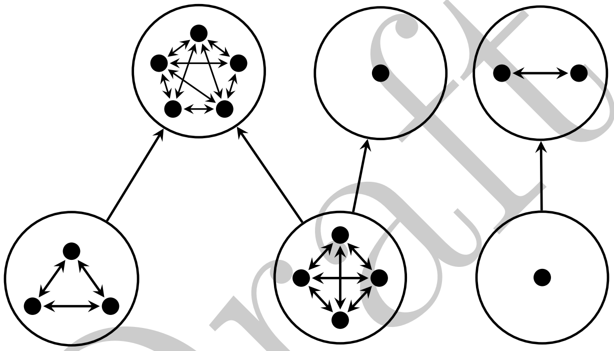


Figure 5: This is the diagram for a single RPF that is not anchored. We cannot turn this into an absolute probability function.

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 ∞. □

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:

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

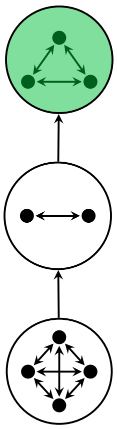


Figure 6: This is a diagram of a totally comparable RPF that is not mutually possible. The mutually possible components form a total order, with the *anchored component* on top.

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 an event is *internally non-comparable*, meaning that it contains outcomes h_1 and h_2 where $P(h_1, h_2) = *$. Perhaps there are interesting things we can say about such an events, but here we will constrain ourselves to totally comparable RPFs in order to avoid such questions.

Definition 6.1. Let P be a totally comparable 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 r, so long as the result isn't indeterminate.

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

Because we no longer have access to absolute probability, the best we can do is measure it relative to a *reference outcome* r. 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 statement 3 if they are non-empty.

Proof. Lemma 5.3 states that all totally comparable RPFs have anchor outcomes, and therefore (by the same argument) every event must contain outcomes that are anchors internally for that event. 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 r in statement 3 yield constraints on P, then they must be equal.
(ii) The constraint in statement 3 does 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 $P(r_1, r_2) = 0$.

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, equation 4 must also hold when $P(r_1, r_2) = \infty$.

So now we can assume that $P(r_1, r_2) \notin \{0, \infty\}$. Multiply the left hand side of equation 4 by $1 = \frac{P(r_1,r_2)}{P(r_1,r_2)}$ and distribute 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 follow from the fact that statement 3 is a ratio with identical expressions 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. *If events e_1 and e_2 are not both empty, the following formula for calculating the relative probability of events is true:*

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

Proof. Find a suitable reference outcome h and multiply by $1 = \frac{P(h_1,r)}{P(h_1,r)}$.

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

Since both $P(e_1, e_2)$ and the formula above match the same thing which is not $*$ for appropriate reference r, 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 relative 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 sum $\sum_{h_2 \in e} P(h_2, h)$ cannot itself be zero because $P(h, h)$ is one of its 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 a unique outcome space. Let P_k measure relative probability on outcome space Ω_k, so that $P_k : \Omega_k \times \Omega_k \to \mathbb{M}^*$.

We can combine all of these relative probability functions together with a top level probability function $P_\top$[10] with outcome space $\Omega_\top = \{\Omega_0, \Omega_1, ...\Omega_{K-1}\}$. The outcome space is heirarchical as shown in figure 7.

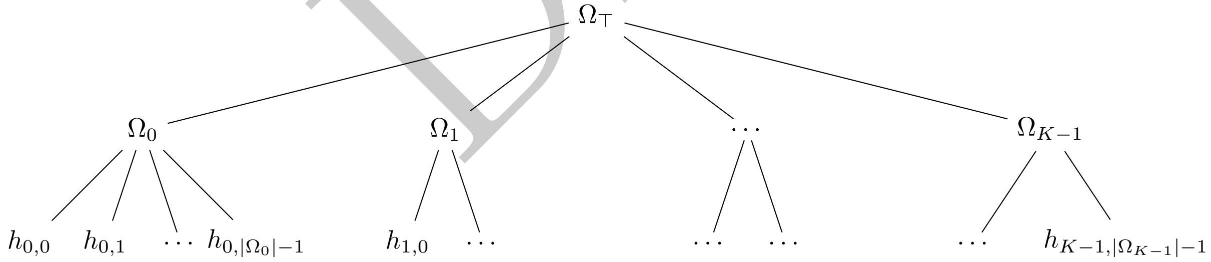


Figure 7: A tree diagram for a set of RPFs being composed by a top-level RPF.

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

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

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

[10] Pronounced "P-Top".

- 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 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 now look at two inputs are from different components.

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 any two outcomes in the same component are always going to be comparable in the overall RPF. 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 sum in the denominator cannot be zero since $P(k_1, k_1) = 1$ will be one of its terms.

If the term $P_{k_1}(h_{k_1,i}) = 0$, 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. Bayesian inference on RPFs is the process of updating these beliefs given new data. The initial RPF is called the *prior* and the updated RPF is called the *posterior*. In many applications, the work is not finished once a formula for the posterior is computed. The practitioner might then want to search the hypothesis space Ω for a model that is either optimal (*maximum a posteriori*) or very good with respect to the posterior distribution. They might also wish to randomly sample one or more values $h \in \Omega$ weighted according to the posterior.

Almost all of these sampling and search algorithms rely exclusively on an iterative algorithm using the relative probability of a current state and certain nearby states[11]. Because relative probability simplifies bayes rule and is ideal for many sampling and selection algorithms, statisticians and engineers should consider using the RPF framework for these purposes.

[11] Examples include hill climbing, simulated annealing, Newton-Raphson, Markov Chain Monte Carlo, and the No U-Turn sampler. The role of these algorithms in supervised machine learning has been previously discussed by Local Maximum Labs in “Sampling Bias Correction for Supervised Machine Learning[2]”

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 two hypotheses and the ratio of their posterior probabilities.

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

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

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

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

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:

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

Let P_D represent the likelihood ratio of the different hypotheses. 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 now as follows: $\frac{P(D|h_1)}{P(D|h_2)} \rightarrow P_D(h_1, h_2)$

These substitution create a bayes rule for relative probability:

$$P(h_1, h_2|D) = P_D(h_1, h_2) P(h_1, h_2) \tag{10}$$

Bayesian inference is now 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 because it 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:

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

□

The following theorems drive home the preference for totally mutually comparable functions for bayesian inference, as pegging one outcome as impossible with respect to another would be a permanent belief. These situations also represent common error modes in bayesian computation.

Theorem 8.2. *If two outcomes are uncomparable in a prior distribution, they will be uncomparable in the posterior distribution. 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. *If an outcome is impossible with respect to another outcome in the posterior distribution, it will either remain impossible or become uncomparable in the posterior. In other words, if $P(h_1, h_2) = 0$, then $P(h_1, h_2|D) \in 0, *$.*

Proof. $P(h_1, h_2|D) = P_D(h_1, h_2)P(h_1, h_2) = P_D(h_1, h_2) \cdot 0$. This finally term would normally simplify to 0, but 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 statistical 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[12]. We recieve the same message several times for redundancy, and we count c_k as the number of times the message was recieved as k.

The indicator function can be used to get the absolute probability of recieving h_1 given that the real message was h_2.

$$P(\text{recieved } h_1|\text{message } h_2) = p[h_1 = h_2] + \frac{1-p}{K}$$

We then use this to construct an RPF for the liklihood ratio if we recieve a single message, k.

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

[12] We could still have gotten lucky and recived the correct value

If we recieve multiple messages in the count vector c, we get the following likelihood formula:

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

We should only care about terms where $k \in \{h_1, h_2\}$ because otherwise the term becomes $\frac{1-p}{1-p} = 1$. We will also assume $h_1 \neq h_2$:

$$\begin{aligned} P_c(h_1, h_2) &= \left(\frac{pK[h_1 = h_1] + 1 - p}{pK[h_2 = h_1] + 1 - p} \right)^{c_{h_1}} \left(\frac{pK[h_1 = h_2] + 1 - p}{pK[h_2 = h_2] + 1 - p} \right)^{c_{h_2}} \\ &= \left(\frac{pK + 1 - p}{1 - p} \right)^{c_{h_1}} \left(\frac{1 - p}{pK + 1 - p} \right)^{c_{h_2}} = \left(1 + \frac{pK}{1 - p} \right)^{c_{h_1} - c_{h_2}} \end{aligned} \tag{11}$$

Because the prior is uniform, the posterior is just equal to the likelihood.

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

We now have an insight: 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 (which may not be required). Using a different prior would be as easy as appending an additional term to the formula for $P(h_1, h_2 | c)$.

8.2 Digital Representation

Even if an inference problem starts with a mutually possible distribution, it could end up anywhere in the RPF chart in figure 2. How can we represent an RPF digitally in a data structure that could account for all of these various possibilities?

The analysis in section 5.4 on mutual possibility classes provides a good framework for this. Every outcome is a member of a mutual possibility class, and has a relative probability within that class.

In addition, we need maintain a partial order of all the mutual possibility classes. The data structure for this purpose could be as simple as listing the outbound edges for each class (illustrated by the graphs in figures 4, 5, and 6). There needs to be a method to compare two classes, and the comparison will return one of four values: greater, lesser, equal, or uncomparable - or in numerical terms $\{\infty, 0, 1, *\}$.

Each comparison might require a traversal of the mutual possibility graph which could get expensive in certain situations. The choice of data structure for partial orders comes with tradeoffs. For small outcome spaces, this question is negligible and graph traversal will usually be adequate. When bayesian inference is performed, the mutual possibility classes will sometimes need to be split up - either vertically as some items in a class become impossible with respect to others - or horizonally as they become uncomparable. The data structure should be able to account for this as well.

Let Ω be the outcome space, and let C be the set of mutually possible classes on Ω with respect to the RPF. We maintain three functions

- $\alpha : \Omega \to C$ assigns outcomes to possibility classes.

- $\ell : \Omega \to (-\infty, \infty)$ provides a log value for an outcome within its mutual possibility class. This allows us to use the full range of floating point numbers available on our machine.
- $Q : C \times C \to \{\infty, 0, 1, *\}$ compares two probability classes.

From these, we can compute the value of an RPF with the following formula:

$$P(h_1, h_2) = Q(\alpha(h_1), \alpha(h_2)) \cdot e^{\ell(h_1) - \ell(h_2)}$$

The simplest RPF to represent in this form is the uniform RPF. In this case, there is a single possibility class so $C = \{0\}$, and ℓ is constant so it can be set as $\ell(h) = 0$. No data is needed for Q because it has an identity rule where $Q(0, 0) = 1$. In fact, Q is itself an RPF on C where instead of providing values in $\mathbb{M}^*$ it provides values in the subspace $\{\infty, 0, 1, *\}$.

9 Topology and Limits in Relative Probability Space

Mathematics can be used to model the real world even through seemingly impossible ideas. For example, we might believe that a natural process can only repeat a finite number of times and that this is just a physical limitation of our universe. And yet at the same time, we will still speak of "infinite iterations" in order to get a bound or estimate on what that system looks 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 result will also be a totally comparable RPF.

This effort caps off a significant argument in favor of relative probability. RPFs hold on to certain pieces information under the limit operation, while absolute probability does not.

Some background in topology[13] required for this section.

9.1 Relative Probability Spaces

Definition 9.1. RPF$^*(K)$ is the set of relative probability functions of size K (where $\Omega = \{0, 1, ..., K-1\}$). Likewise RPF$(K)$ is the set of all totally comparable RPFs of size K.

Because the set of absolute distributions with $|\Omega| = K$ is embedded in $\mathbb{R}^K$ as seen in figure 1, its topological properties are well understood. The simplex is closed, bounded, and compact.

For relative probability distributions, there is no obvious way to embed it into K-dimensional euclidean space[14]. The relative probability space is more complicated, because at the corners and edges of the simplex lurk entire subspaces where zero-probability outcomes are still being compared in different configurations.

Fortunately, the set of all RPFs can still be embedded into a much larger euclidean space. Any $P \in$ RPF(K) is a function of type $\Omega \times \Omega \to \mathbb{M}$ that satisfies the fundamental properties. Therefore RPF(K) can at least be embedded into $\mathbb{M}^{K^2}$.

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

[14] Though it may be possible! See section 10.3

A topology can be defined by as *basis of open sets*, and for euclidean space (and metric spaces more generally) the basis is the set of open balls on $\mathbb{R}^n$.

Definition 9.2. An open ball of size ϵ around point $x \in \mathbb{R}^n$ is the set of all points y such that $|x-y| < \epsilon$.

Open balls do not work as a basis on the set $\mathbb{M}$ because an open ball around ∞ would contain only ∞ - when in reality we want an open set around ∞ to also contain some interval $(x, \infty]$. This is remedied by using the following transformation, which we take to be continuous (topology conserving) between $\mathbb{M}$ and $[0,1]$.

Definition 9.3. The *inverse odds transform* is the function $\text{odds}^{-1} : \mathbb{M} \to [0,1]$ with $\text{odds}^{-1}(0) = 0$ and $\text{odds}^{-1}(\infty) = 1$ defined by

$$\text{odds}^{-1}(x) = \frac{x}{x+1}$$

The inverse odds transformation establishes $\mathbb{M}$ as topologically equivalent to the closed interval $[0,1]$. Now $\text{RPF}(K)$ can be embedded into a bounded region of euclidean space, namely $[0,1]^{K^2}$.

There are other ways to define a topology on RPF^K. One in particular using compositions of open patches was discarded due to its complexity during the course of this research[15].

9.2 Limit Points and Compactness

We now close with an argument for the closure of $\text{RPF}^*(K)$ under limits.

Definition 9.4. A *limit point* of a set A is a point x such that any open set containing x also contains points in A. Equivalently for euclidean space, any open ball containing x also intersects with A.

Theorem 9.1. *Let P be a limit point of RPF(K). Then P satisfies all of the fundamental axioms and is therefore a member of RPF(K).*

Proof. Let P be a limit point of $\text{RPF}(K)$. We can show that for each fundamental axiom, if P doesn't satisfy the axiom then some open ball around P also doesn't satisfy the axiom.

Identity: If P doesn't satisfy the identity axiom, then $P(h,h) \neq 1$ for some outcome h. The inverse odds transform maps the value 1 to $\frac{1}{2}$, so a choice of ϵ less than $|\text{odds}^{-1}(P(h,h)) - \frac{1}{2}|$ will contain only elements P' where $P'(h,h) \neq 1$.

Inverse: A similar argument applies here if for some pair h_1 and h_2, $P(h_1,h_2) \neq P(h_1,h_2)$. An ϵ can be selected that is small enough so that $P(h_1,h_2)$ is never equal to $P(h_1,h_2)$

Composition: Suppose $P(a,c) :\ncong P(a,b) \cdot P(b,c)$. This means that the constraint cannot be $*$ by lemma 2.2, so by lemma 2.1 we can conclude that $P(a,c) \neq P(a,b) \cdot P(b,c)$. But then again, the same argument from identity and inverse apply; for some sufficiently small ϵ-ball around P, the composition axiom will still always be false.

This means that only functions P which satisfy the fundamental axioms can be limit points to $\text{RPF}(K)$, and therefore $\text{RPF}(K)$ contains all its limit points. □

[15] Its usefulness is dubious, but it will be posted separately as an addendum note in case that assessment turns out to be incorrect

As a bonus, we can prove that RPF(K) is compact thanks to the Heine-Borel theorem, stated as follows (wording from Bradley [4]).

Theorem 9.2 (Heine-Borel Theorem)**.** *A subset of $\mathbb{R}^n$ is compact if and onely if it is closed and bounded.*

Theorem 9.3. *RPF(K) is compact.*

Proof. Theorem 9.1 states that RPF(K) is closed, and by the odds transform it is also bounded. By the Heine-Borel theorem, it is compact. □

10 Future Work

10.1 Expansions to infinite spaces

The obvious extention to this work is to expand relative probability to a generalized space which may be infinite, and thus capture all of the variety of probability distributions that one might wish to study and apply. This would start by modifying statement 3 to ask for an additive property. The relative probability function would then become a *relative probability distribution.*

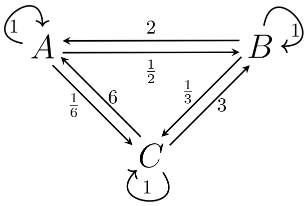
This raises certain question which - while decisions have been made in prior work and certainly in measure theory - should be open to discussion.

1. We may not need to keep countable additivity as set forth in the Kolmogorov axioms; we can relax this to allow for a fair countable lottery (also known as a De Finetti Lottery[17]). RPFs would provide a great way to analyze the fair countable lottery as a uniform distribution and this should be exploited!
2. If we derive a notion of probability density, then can these densities at a particular pair of events be used to compare the relative probability of those events? What specific properties of the relative probability distribution are required to make this work?

It appears possible to use these ideas to create a unified version of the Hausdorff measure - which finds the size of an object given its dimention. Instead of considering it to be multiple measures - we can have a single measure where bounded sets of equal dimention are mutually possible, and smaller-dimensional objects are always mutually impossible with respect to a larger dimensional objects.

10.2 Relationship to Category Theory

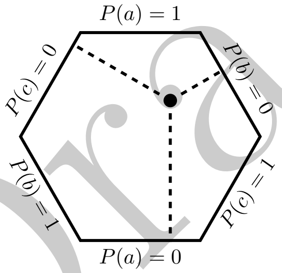
Category theorists will recognize that an RPF describes a category. Specifically, and RPF describes something called a *thin category* where any pair of objects have at most one morphism connecting them (per direction). The relative probability axioms can analyzed and approached through the lens of category theory in order to learn more about them.



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

10.3 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 but it should still be possible given that RPF^K is a (K-1)-dimensional set that can be embedded into $[0,1]^{K^2}$. 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 simplices, this appears to become exceedingly unweidly unless some simplifying trick is developed. If it is successfully done, then a new, cleaner, representation for memebers of RPF(K) falls into place.

10.4 Dirichlet Distributions over $RPF(K)$

The Dirichlet distribution of dimention K is a probability distribution **over** the K-dimensional simplex. This distribution's PDF may assign 0 to various edges and corners of the simplex when it's parameters are large. When it's parameters approach zero, it assigns zero probability to the interior of the simplex and infinite probability to the boundary.

The Dirichlet distribution should be re-analyzed as a (continuous) relative probability measure over the space of RPF(K). This allows the set of possible dirichlet distributions to take on its limit values.

The relative version of the dirichlet is given as follows. The there are K magnitude parameters represented by $\alpha = (\alpha_0, \alpha_1, ..., \alpha_{K-1})$ with the type of α being $\mathbb{M}^K$. The relative probability compares two RPFs, P_1 and

P_2. It makes use of a reference outcome r, similar to the event comparison in definition 6.1. The simplifying powers of relative probability are on full display.

$$\mathcal{D}_\alpha(P_1, P_2) = \prod_{k=0}^{K-1} \left(\frac{P_1(k, r)}{P_2(k, r)} \right)^{\alpha_k - 1}$$

The Dirichlet is already extremely useful to machine learning in areas such as topic modeling[16] and prior distribution generation[1]. It represents a "fuzzier" version of the categorical distribution for applications where we do not want to rely on knowing the categorical exactly at the expense of computational complexity. If the Dirichlet were allowed to include its limit values, then it would be a valuable tool for feature selection by telling us where to rely on a pure categorical parameters, where to hedge on the (fuzzier) dirichlet.

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] Rényi, A. (1955). On a new axiomatic theory of probability. Acta Mathematica Hungarica, 6(3-4), 285-335.

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

[12] Heinemann, F. (1995). Closing Economic Models by Information Assumptions. Inst. für Volkswirtschaftslehre u. Statistik d. Univ. Mannheim.

[13] November, D. D. (2018). Zero Probability.

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

[15] Kohlberg, E., & Reny, P. J. (1997). Independence on relative probability spaces and consistent assessments in game trees. Journal of Economic Theory, 75(2), 280-313.

[16] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. the Journal of machine Learning research, 3, 993-1022.

[17] Finetti, B. D. (1974). Theory of probability, vol. 1 and vol. 2.

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.