Relative Probability on Finite Sample Spaces
SUBTITLE HERE

Max Sklar
Local Maximum Labs
DATE HERE

Abstract

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

Draft

1 Introduction

It may surprise many that the mathematical foundations of probability theory are still very much open to debate.

Since Kolmogorov published the standard axioms for probability theory in 1933, there have been calls to relax some of those assumptions and rules for various applications. In the paper Kolmogorov's Axiomatisation and Its Discontents[5], Lyon lays out these cases and their justification. One of difficult questions is the definition of a condiational probability where the condition has a probability of zero. Informally, this means that we wish to talk about the probability of an event given a hypothetical situation which will not occur[1]. Related to this question is whether we can talk about the relative probability of two events in a system, even though those two events may have probability zero.

The standard model defines these values as a ratio of one probability to another. As a result, every time the indeterminate form $\frac{0}{0}$ appears, the relative probability must remain undefined. Undeterred by this state of affairs, mathematicians and engineers talk about this type of relative probability all the time. For example, if we consider the 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* or *PDF* values.

Let us take the position that we may model probability in a non-standard way, and we can do so as long as the framework is logically consistent, and the advertised applications match such a proposal[2]. Given the popularity of Bayesian methods as applied to machine learning in recent decades, and given that the methods used to search a hypothesis space in Bayesian inference relies on relative probability[2], we ought to understand whether we can derive a framework for probability that takes the relationships between outcomes and events as fundamental.

[1] There is another very unintuitive feature of probability theory zero probability events do indeed occur all the time, particularly when given a continuous distribution.

[2] Lyon identifies this link to application 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 a person would reasonbly call probability, then it ought to go by a different name.

Not only would this solve the conditional probability question, but it would also give rise to several new concepts and objects to study. In this paper we will construct a theory of relative probability on finite distributions as a starting point for more complexity in future work. As we shall see, even for this simple case there is much work to be done!

We will also look at potential applications of these ideas, including limit distributions, bayesian inference, and even implementation.

These methods have applications in Bayesian statistics, even providing a new formulation of Bayes rule. This new formulation is reflective of current practice.

2 Preliminaries

2.1 Magnitude Space

Let $\mathbb{M} = [0, +\infty]$ be the space of *magnitudes*, which roughly corresponds with our intuition for the concept of size. The magnitude space contains all positive real numbers, 0 and ∞.

The following functions are defined for all m_1, m_2 in $\mathbb{M}$.

$\mathbb{M}$ is closed under addition. $m_1 + m_2$ is the sum of m_1 and m_2.

$$m + \infty = \infty$$

The products $0 \cdot \infty$ and $\infty \cdot 0$ are indeterminate. In all other cases, multiplication on $\mathbb{M}$ is defined.

The multiplicative inverse m^{-1} is the defined on all $\mathbb{M}$. We let $0^{-1} := \infty$ and $\infty^{-1} := 0$, even though it doesn't quite act as a multiplicative inverse in those cases.

Any magnitude can be raised to the power of a real number and still be a magnitude.

$\forall x \in \mathbb{R}, m^x \in \mathbb{M}$.

If we consider the *extended real number line* to contain values at $+\infty$ and $-\infty$, we can continue to define these exponential operations, except for when $m = 1$:

If $m > 1$, then $m^{+\infty} = \infty$ and $m^{-\infty} = 0$ If $m < 1$, then $m^{+\infty} = 0$ and $m^{-\infty} = \infty$

The set of magnitudes is a totally ordered set under $\leq$, meaning that for every two magnitudes either $m_1 \leq m_2$ or $m_2 \leq m_1$, and if both are true then $m_1 = m_2$

2.2 The Wildcard Element

$$\mathbb{M}^* = \mathbb{M} \cup \{*\}$$

Draft

Let the *magnitude-wildcard space* $\mathbb{M}^*$ be defined as the set of magnitudes along with a *wildcard element*, $*$. The wildcard element corresponds to several different ideas depending on the context:

- The *wildcard pattern* used in pattern matching and regular expressions in type theory and computer science
- The indeterminate form $\frac{0}{0}$ in arithmatic.
- The standard *NaN*, or *Not a Number*[3] value in the IEEE standard for floating point arithmetic.

We define the following properties on $*$:

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

(ii) $* + m = *$

(iii) $* \cdot m = *$

Let the **matching relation**[4] $:\cong$ be a binary relation on $\mathbb{M}^*$. A value m_1 is matched by m_2 whenever m_2 is the wildcard, or the two values are equal.

$$m_1 :\cong m_2 \iff [m_2 = *] \vee [m_1 = m_1]$$

From this definition, several important properties quickly follow.

(i) If a magnitude matches a non-wildcard element, then the two values are equal.

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

(ii) Every element is matched by the wildcard element.

$$m :\cong *$$

(iii) The wildcard element is matched only by itself.

$$* :\cong m \implies m = *$$

The matching relation ask if the element on the left hand side could be represented by the element on the right. The wildcard element represents every single value, but it cannot be represented by any specific value. To put it another way, the wildcard element represents a loss of information about a value which can never be recovered.[5]

The matching relation is reflexive, but in general is not symmetric.

Theorem: The matching relation is transitive. In other words, for all m_1, m_2, m_3 in $\mathbb{M}$, if $m_1 :\cong m_2$ and $m_2 :\cong m_3$, then $m_1 :\cong m_3$.

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

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

[5] Hajek[6] calls these matching relations contraints, in that they may or may not bind the right hand side to a specific value.

Draft

Proof: Assume that $m_1 :\cong m_2$ and $m_2 :\cong m_3$. If none of these values are the wildcards, then by property (i), they are all equal and $m_1 :\cong m_3$. If $m_1 = *$ then by property (iii), $m_2 = *$ and finally $m_3 = *$ so the theorem holds. If $m_2 = *$ then $m_3 = *$ and $m_1 :\cong m_3$ by property (ii). And of course if $m_3 = *$ alone, then by property (ii), m_1 is still matched by m_3.

Theorem: The matching relation preserves multiplication.

$$\forall a, b, a', b' \in \mathbb{M}^*, a :\cong a' \wedge b :\cong b' \Rightarrow ab :\cong a'b' \tag{1}$$

Proof: Let $a, b, a', b' \in \mathbb{M}^*$, and let $a :\cong a'$ and $b :\cong b'$. If either a' or b' are a wildcard, then $a'b'$ is also a wildcard by definition of wildcard multiplication and the theorem holds. If a' and b' are not wildcards, then $a = a'$ and $b = b'$, making the statement obvious.

3 Categorical Distribution

3.1 Definition

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

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 can be embedded in $\mathbb{R}^K$ as a (K-1)-dimentional object called a (K-1)-simplex. 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).

Because the set of categorical distributions is embedded in Euclidean space $\mathbb{R}^K$, its topological properties are well understood. We take the topology of the simplex to be the sub-topology from $\mathbb{R}^K$. The simplex is closed, bounded, and compact. Practically, this means that any sequnce of points on the simple will converge to one or more points on the simplex allowing both pure and applied practicioners to talk about limit and boundary conditions.

Draft

3.2 Events

An *event* is a set of outcomes. We define $\mathcal{F}$ as the space of all possible events. $\mathcal{F}$ is the power set[6] of Ω, meaning that $\mathcal{F} = \mathcal{P}(\Omega)$, and for any event $e \in \mathcal{F}$, $e \subseteq \Omega$.

In the previous section, we defined the probability of individual outcomes. We can now define the probability function on events instead of outcomes. For finite distributions, this is simple. 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 $P(\{h\}) = P(h)$.

[6] In general, $\mathcal{F}$ is not the entire power set of Ω but typically is when Ω is finite.

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

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

3.3 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.

$$P(e_1, e_2) = \frac{P(e_1)}{P(e_2)} \tag{2}$$

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:

$$\forall e_1, e_2, e_3 \in \mathcal{F}, P(e_1, e_3) :\cong P(e_1, e_2) \cdot P(e_2, e_3)$$

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

4 Relative Categorical Probability Functions

In section 3.3, 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 Relative Probability Functions on Outcomes

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

Draft

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

The expression $P(h_1, h_2)$ can be read as the probability of h_1 relative to h_2. A relative probability function is valid if and only if it obeys *the fundamental properties of relative probability* as follows.

The *reflexive property of relative probability* states that the relative probability of every element relative to itself is one.

$$P(h, h) = 1 \tag{3}$$

[ADD SYMMETRIC]

The *transitive property of relative probability* states that the relative probabilities of outcomes must make sense when compared to a third outcome.

$$P(h_1, h_3) :\cong P(h_1, h_2) \cdot P(h_2, h_3) \tag{4}$$

4.2 From Outcomes to Events

Our next task is to upgrade P to operate on the event level. In definition 2, the relative probability of two events were set as the ratio of their absolute probabilities. Because we no longer have access to absolute probability, the best we can do is measure relative probability to a given event h^*. Because this ratio might be indeterminate, we use the matching relation instead of equality.

For all $h^* \in \Omega$,

$$P(e_1, e_2) :\cong \frac{\sum_{h_1 \in e_1} P(h_1, h^*)}{\sum_{h_2 \in e_2} P(h_2, h^*)} \tag{5}$$

In addition, we require that P obeys the fundamental properties of relative probability given in equations 3 and 4. Finally, if a relative probability is not constrained by any of these requirements, then it remains a wildcard.

These requirements seem reasonable, but how can we know for sure that they consitute 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 5 yield non-wildcard constraints on P, then they must be equal.

(ii) If $e_1 = e_2$, then the right hand side of statement 5 for $P(e_1, e_2)$ will always yield 1 or $*$.

(iii) The transitive property does not place any new constraints on P not already placed by the reflexive property or statement 5. With P now fully defined, it obeys the transitive property.

Draft

Proof of (i)

Let h_1^* and h_2^* be two distinct values for h_* in 5, and neither causes the ratio to be a wildcard. Then we want to check that

$$\frac{\sum_{h_1 \in e_1} P(h_1, h_1^*)}{\sum_{h_2 \in e_2} P(h_2, h_1^*)} = \frac{\sum_{h_1 \in e_1} P(h_1, h_2^*)}{\sum_{h_2 \in e_2} P(h_2, h_2^*)} \tag{6}$$

None of the terms above can be wildcards, because it would cause the whole expression to be a wildcard. The key to this argument is in looking at the value of $P(h_1^*, h_2^*)$. Fortunately, a few potential values can be eliminated.

Suppose $P(h_1^*, h_2^*) = *$. Then for all h, $P(h_1^*, h_2^*) = * :\cong P(h_1^*, h) \cdot P(h, h_2^*)$. By matching property (iii), this means that $P(h_1^*, h) \cdot P(h, h_2^*) = *$, and therefore at least one of those terms is $*$. But looking back on equation 6, this means that if there are any terms at all in these sums, at least one of them must be $*$, contradicting our assumption. And if there happen to be no terms in this equation then both sides are $\frac{0}{0} = *$. So $P(h_1^*, h_2^*) \neq *$

Perhaps $P(h_1^*, h_2^*) = 0$. Then, every term from equation 6 of the form $P(h_1, h_2^*)$, we have $P(h_1, h_2^*) :\cong P(h_1, h_1^*) \cdot P(h_1^*, h_2^*) = P(h_1, h_1^*) \cdot 0$. Then either $P(h_1, h_2^*) = 0$ or $P(h_1, h_1^*) = \infty$. More generally, equation 6, this means that every term in equation 6 is either 0 or ∞ and its analogous term on the opposite side is the inverse.

Looking at the sums as a whole, each sum is made up either entirely of 0s or contains an ∞ in which case the sum is ∞.

[GOTTA FILL THIS IN]

By an analogous argument $P(h_1^*, h_2^*) \neq \infty$

So now we can assume that $P(h_1^*, h_2^*) \notin \{0, \infty, *\}$. The allows us to multiply the left hand side of equation 6 by $1 = \frac{P(h_1^*, h_2^*)}{P(h_1^*, h_2^*)}$ and distribute into the sum to get:

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

Note that the simplification with the transitive rule needs to be justified in the wildcard space. We can be sure that none of the terms $P(h_1, h_1^*)$ or $P(2_1, h_1^*)$ are $*$ because that would make the entire expression $*$ which contradicts our assumption. Because $P(h_1^*, h_2^*)$ is a magnitude between 0 and ∞, we can be sure that the transitive simplification holds.

Proof of (ii)

$$P(e_1, e_1) :\cong \frac{\sum_{h_1 \in e_1} P(h_1, h^*)}{\sum_{h_1 \in h_2} P(h_1, h^*)}$$

The right hand side, being the ratio of two identical terms, is either 1 or *ast* in the case that the sums are 0 or ∞. Therefore, there cannot be a contradiction to $P(e_1, e_1) = 1$

Draft

Proof of (iii)

The fundamental property on the event level is $P(e_1, e_3) :\cong P(e_1, e_2) \cdot P(e_2, e_3)$

FILL IN

4.3 Formulas for the Event Level Function

Theorem: Given events e_1 and e_2 where $e_1 \cup e_2 \neq \varnothing$, 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: FILL IN

We then derive the absolute probability function as

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

Theorem: For all RPF P and events e_1 and e_2, $P(e_1, e_2) = P(e_2, e_1)^{-1}$.

Proof: Using the transitive property, $P(e_1, e_1) = 1 :\cong P(e_1, e_2) \cdot P(e_2, e_1)$.

FILL IN

Definition: Events e_1 and e_2 are *comparable* if $P(e_1, e_2) \neq *$.

Draft

4.4 Degenerate Cases

Suppose that $K = 0$. Then Ω is empty, and there are no outcomes. Surprisingly, there is still an RPF because of the event represented by the empty set $\varnothing$. In this case $P(\varnothing, \varnothing) = 1$ is the only RPF. This is a mathematical by product of our definition, but an interesting comparison to the case of absolute distributions where such a function does not exist (because with no outcomes, they cannot sum to 1).

Now we let $K = 1$, and $\Omega = h$. There is still only a single, trivial RPF P, where $P(h, h) = 1$, and $P(\varnothing, h) = 0$. This matches the absolute case where the probability of the single outcome must be 1.

4.5 Matching and Equality

Let P_1 and P_2 be relative probability functions.

Definition: $P_1 :\cong P_2$ if and only if all of relative probabilities of P_1 are matched by those of P_2.

$$\forall e_1, e_2 \in \mathcal{F}, P_1(e_1, e_2) :\cong P_2(e_1, e_2) \tag{7}$$

A relative probability function is *totally comparable* if every pair of events are comparable. [CITE DEFINITION USING amsthm]

Theorem: Let P_1 and P_2 be relative probability functions on Ω where $P_1 :\cong P_2$ and P_2 is totally comparable. Then, P_1 and P_2 must be equal.

Proof: Assume that $P_1 :\cong P_2$ and choose any two events, e_1 and e_2. By definition, it directly follows that $P_1(e_1, e_2) :\cong P_2(e_1, e_2)$. Because P_2 is totally comparable, $P_2(e_1, e_2) \neq *$. By matching property (iii), this means that $* :\ncong P_2(e_1, e_2)$. This in turn means, that $P_1(e_1, e_2) \neq *$ because if it were then it would not be matched by $P_2(e_1, e_2)$. Since neither expression is a wildcard, using matching property (i) we conclude that $P_1(e_1, e_2) = P_2(e_1, e_2)$. Therefore, these two RPFs are equal on all possible inputs.

Theorem: Every RPF is matched by an absolute probability function

Proof: FILL IN

Theorem: An absolute probability function is totally comparable if and only if at most 1 outcome is assigned 0 probability.

Proof: Let P be an (absolulte) probability function, with h_1 and h_2 being two outcomes. If $P(h_1) = P(h_2) = 0$, then the relative function (CITE) $P(h_1, h_2) = \frac{0}{0} = *$, and thus P is not totally comparable. 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. Therefore, P is comparable if only 1 outcome is assigned 0 probability.

4.6 Possibility Classes

Events e_1 and e_2 are *mutually possible* if they are comparable and $0 < P(e_1, e_2) < \infty$.

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

Proof: For reflexive, $P(e_1, e_1) = 1$, so every event is comparable to itself.

For symmetric, $P(e_1, e_2) = P(e_2, e_1)^{-1}$, which means that each can be in $\{0, \infty, *\}$ if and only if the other one is as well.

For transitive, we use the fundamental property for RPFs which states that $P(e_1, e_3) :\cong P(e_1, e_2) \cdot P(e_2, e_3)$. If the last 2 values are positive real numbers, then their product is also a positive real number and equal to $P(e_1, e_3)$.

The event e_1 is impossible with respect to e_2 if $P(e_1, e_2) = 0$. The event e_1 is possible with respect to e_2 if they are comparable and $P(e_1, e_2) > 0$

Theorem: The relationship of being possible is a *preorder*, or in other words both reflexive and transitive.

Proof: It must be reflexive because $P(e, e) = 1$. If $P(e_1, e_2) > 0$ and $P(e_2, e_3) > 0$ then their product is also greater than zero, and by the fundamental property, equal to $P(e_1, e_3)$ thus making it transitive.

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

Draft

we have a partial order.

The theorems above can be summerized as follows. Given any two events, e_1 and e_2, exactly one these is true.

(i) e_1 and e_2 are mutually possible.

(ii) e_1 is impossible with respect to e_2.

(iii) e_2 is impossible with respect to e_1.

(iv) e_1 and e_2 are not comparable.

A relative probability function is called *mutually possible* if all of its outcomes[7] are mutually possible. Mutually possible RPFs satisfy *Cromwell's rule* in Bayesian inference, which states that prior beliefs should assign probability zero or one to events[8].

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

Proof: FILL IN

Draft

4.7 Totally Mutually Possible

Definition: A relative probability function is *totally mutually possible* if all outcomes are mutually possible.

Theorem: A totally mutually possible RPF can be represented by an absolute probability function $P : \Omega \to (0, 1)$ which contains no zero-probability events.

Proof: Define the absolute probability of outcome h as $P(h, \Omega)$. By definition of relative probability on events we have for every outcome h^*:

$$P(h) = P(h, \Omega) :\cong \frac{P(h, h_*}{\sum_{h_1 \in \Omega} P(h_1, h^*)}$$

Now let $h^* = h$

$$P(h) :\cong \frac{P(h, h)}{\sum_{h_1 \in \Omega} P(h_1, h)} = \frac{1}{\sum_{h_1 \in \Omega} P(h_1, h)}$$

Because P is totally mutually possible, $P(h_1, h)$ will always be a value $\notin \{0, \infty, *$. Therefore, the sum is the same and each $P(h)$ is non-zero. This means that all of the relative probabilities $P(h_1, h_2)$ will have to equal the ratio $\frac{P(h_1)}{P(h_2}$. Therefore, the absolute probability function contains all the information needed to construct the relative probability function without loss of information.

[7] Note that this applies to individual outcomes and not events. The empty event $e = \{\}$ will be impossible with respect to any outcome.

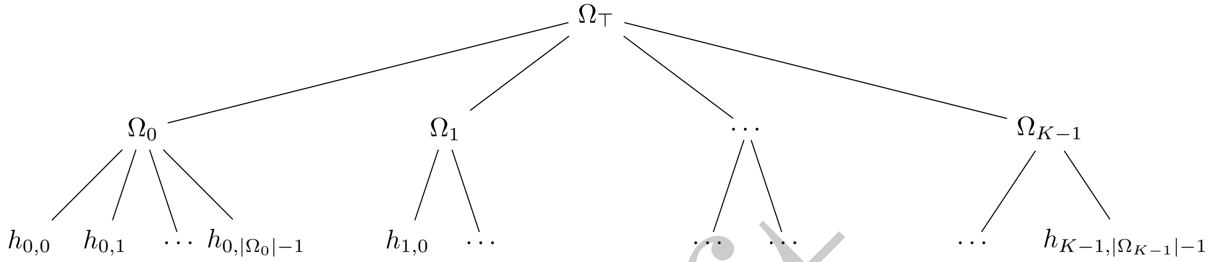
[8] It would be stated differently in continuous space of course.

4.8 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 $P_\top$[9] 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 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:

$$P(h_{k,i}, h_{k,j}) = P_k(h_{k,i}, h_{k,j}) \tag{8}$$

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

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

Note the use of the transitive property to traverse up and down the tree. One could of course imagine this for a tree of many levels, or even each level having variable height.

Theorem: P respects the fundamental property.

Proof:

Reflexive is obvious because an outcome is on the same component as itself, meaning that we can use equation 8 to get $P(h_{k,i}, h_{k,i}) = P_k(h_{k,i}, h_{k,i}) = 1$

[Add Symmetric]

Transitive is also obvious if all of the values are in the same component as it reduces to the transitive property within that component.

Draft

[9] Pronounced "P-Top".

[Add Transitive for different components]

Theorem: 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) There is at most one component with outcomes that are impossible with respect to that component. Equivalently, if $h_{k_1,i}$ and $h_{k_2,j}$ are two components, and $P_{k_1}(h_{k_1,i}) = P_{k_2}(h_{k_2,j}) = 0$, then $k_1 = k_2$. Also equivalently, all components except at most one are totally mutually possible.

Proof: FILL IN (both directions)

5 Topology of Relative Probability Spaces

The idea of a limit distribution requires particularly around the idea of limits. Mathematics can be great at modelling the real world even if it's ideas are theoretically impossible. For example, we might believe that it is impossible for a certain natural process to repeat an infinite number of times, and yet we may still take its value to be infinity in order to get some kind of bound on what that system will look like in the long run. Likewise, it still makes sense to consider a particular outcome in a probabilistic system as certain while maintaining information about the other outcomes in order to calculate the effects of such a limit. [CLEAN UP WORDING OF PARAGRAPH] To this end, we will prove that the space of fully comparable relative distributions is compact.

TODO: Warn people that background in topology is required for this section, and then we can shorten it up! Also, this section can be skipped if not interesting.

One of the benefits of relative probability spaces is their properties with respect to limits. Specifically, if we look at the space of (absolute) categorical distributions on Ω and we allow the probability of one outcome to approach 1, then all of the other probabilities will be forced down to 0 and become incomparable with one another. In the relative probability space, the information about the ratios of probabilities of the other outcomes can be preserved even as a single outcome reaches a probability of 1.

If the space of totally comparable RPFs can be shown to be compact, then we know that relative probabilites of outcomes and events are preserved even as they both approach zero relative to another event.

In order to prove compactness, we first must define a topology on the space of totally comparable RPFs. This means identifying the sets that are open (or intuitively, sets that fully surround all of it's members and don't contain its boundary). This starts with finding a *basis of open sets* from which all other open sets can be constructed through countable unions.

For the absolute probability function, we can use $K-1$-simplex embedded in $\mathbb{R}^K$ to get a topology using the standard Euclidean space. There's no obvious way to embed an RPF into euclidean space, so some background is required.

First note that 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. It can

Draft

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 $\mathbb{R}^2$. [ILLUSTRATION]

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.

Theorem: 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 event level.

Definition: An *interior open patch* of RPF(Ω) 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.

TODO - diagram for interior and exterior open patches

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

1. If $K = 2$, a subset parameterized by an open interval of magnitudes containing 0. $\{P|0 \leq P(h_1, h_2) < b\}$ for some $b \in \mathbb{M}$
2. If $K > 2$, a composition where P_{TOP} is an exterior open patch of size 2, and the component that might be zero relative to the other component - h_1 above - is also an exterior open patch while the second component is an interior patch.

Intuition: 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!

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

Every element of an open patch of RPF(Ω) is totally mutually comparable.

Now let the open patches be the bases for an open set thus defining a topology on the set of totally mutually comparable RPFs of Ω.

Theorem: The topological space of totally comparable functions on an outcome space Ω is *compact*, meaning that for every open cover of it, there is a finite subcover.

Proof - STILL A LOT TO DO:

Draft

Let $K = |\Omega|$. This is going to be 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 the set of open sets is finite, so we're good. (Reference degenerate cases)

Let $h \in \Omega$ be an outcome. The space of totally comparable functions on Ω can be split into 2 regions: one where $P(h, \Omega) = 0$ and one where $P(h_1, \Omega) > 0$.

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

5.1 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: 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

Draft

6 Bayesian Inference on Relative Distributions

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

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

$$P(h|D) = \frac{P(D|h) \cdot P(h)}{P(D)}$$

Now we are going to consider relative probability by looking at the ratio between two hypotheses.

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

Next make the following subsitutions:

For the ratio of absolute distributions, subsitute the relative distribution: $\frac{P(h_1)}{P(h_2)} \rightarrow P(h_1, h_2)$

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

The liklihood ratio: $\frac{P(D|h_1)}{P(D|h_2)} \rightarrow P(D|h_1, h_2)$

Now we get bayes rule for relative probability:

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

The likelihood ratio $L(D|h_1, h_2)$ is essentially a description of how the different hypotheses rate the likelihood of data. It is also a relative probability in its own right, and must obey the fundamental properties (cite).

Therefore, the act of bayesian inference is an element-by-element multiplication of two different RPFs $L(D|h_1, h_2)$ and $P(h_1, h_2)$. We show that the product of two RPFs obeys the fundamental properties.

Theorem: 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 properties.

Proof: Use the multiplication property of the matching relation in equation 1.

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

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

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

Theorem: 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: $P(h_1, h_2|D) = L(D|h_1, h_2)P(h_1, h_2) = L(D|h_1, h_2) \cdot * = *$

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

Draft

7 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.

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 mutually comparable? Algorithm: Look for a linear list of tiers.

Draft

8 Future Work

8.1 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.

8.2 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

8.3 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

8.4 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.

Draft

Appendices

A Is an Appendix Needed

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.

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.