Non-Deterministic Distributions

Christopher Landauer
Aerospace Integration Science Center, The Aerospace Corporation,
P.O. Box 92957, Los Angeles, California 90009-2957
E-mail: cal@aero.org

Abstract

Most models of uncertainty change a precise value in some formal space to some other kind of value, function or a set, in another formal space. In this paper, we define a new representation called non-deterministic distributions, which allow us to integrate ignorance with probability models, and prove several theorems relating them to information theory and denotational semantics. The model integration problem is very difficult in general; we, like many others, have solved it only in very special cases. Our approach to integrating ignorance with probability, using non-deterministic distributions, shows a way to integrate ignorance with other uncertainty measures.

1 Introduction

An uncertainty model is a knowledge representation, either of the state of being of an external phenomenon, or of the state of our knowledge about that phenomenon. Most models of uncertainty are precise models of imprecise reasoning. Instead of using a precise or specific value in some formal space, they use some other kind of value, either a function or a set, in another formal space. They start with a domain of discourse that is a set, and generally require estimation of many more numerical parameters to characterize the uncertainty than there is usually data to support. Each approach has developed a collection of heuristics and special cases to reduce the estimation load.

There are many different meanings of uncertainty, and in many models these are combined in unjustified or ad hoc ways. Hence, a rulebase may include rules that deal with a human’s estimate of likelihoods (“subjective probabilities”), rules that characterize phenomena that have complex dynamics (as in the chaotic dynamics of certain weather phenomena), rules that characterize the results of measurement error, and rules that suggest strategies for how to deal with intermediate actions until the next data sample will resolve current unknowns into known quantities. Often any source of variation, any imprecision, any degree of vagueness or ambiguity is treated in a probabilistic fashion. We call all of these models “uncertainty” models, though there are many other terms in use in special cases.

Uncertainty models are always models, that is, they are representations of some phenomenon in a formal space (whether the phenomenon itself is in a formal space or not) [2]. Therefore, a model of uncertainty is not well-defined without a definition of that formal space, and thus an explicit list of the entities that can be used for representation, and the operations that are available on them. These operations determine what questions that the model can answer, so the modeler can choose a model according to which questions it is expected to answer. A common failure in the use of uncertainty models is to ignore the operations, and assume that the only important part of a formal space is its elements. In particular, it is a common error to presume that real numbers between 0 and 1 are probabilities.

In this paper, we introduce a novel but ad hoc integration of ignorance models with probability distributions (so we can avoid the law of indifference that equates ignorance with uniform distributions), and concentrate on some of the mathematical properties that make it quite natural, in our opinion. We define non-deterministic distributions, our new method of integrating ignorance with probability, so that one can specify explicitly not only what one knows about a probability, but also what one does not know.

The model integration problem is very difficult in general; we, like many others, have solved it in very special cases [2] [8], and are working towards more generality [10]. We believe that our approach to integrating ignorance with probability, using non-deterministic distributions, shows a way to integrate ignorance with many other uncertainty measures. With our wrapping integration infrastructure, described in detail elsewhere [10] [11], we have a method for approaching the general problem, and a way to implement testbeds to study any method of integration.
2 Non-Deterministic Distributions

This section describes a generalization of probability distributions over finite sets from elements of those sets to subsets. We make no peculiar philosophical assumptions beyond the ones required to justify using probability distributions in the first place.

Our main point of departure from the standard Bayesian methodology is that we distinguish between not knowing anything about a probability distribution and knowing that the distribution is uniform, and provide mechanisms for measuring the distinction. We define a “less specific” ordering and relate it to the partial orderings of denotational semantics, and also provide a “best guess” probability distribution for situations that do require a probability distribution without providing enough information to define one.

We make a formal distinction between knowledge of uniform probability and no knowledge, best typified by the distinction between a known fair two sided coin, for which the probabilities of head and tail are each one half, and an unknown two sided coin, for which we have no guarantees of behavior at all, beyond the two known alternatives. For the fair coin, we have some uncertainty in the individual outcomes, but some guarantees of the asymptotic behavior.

Many studies of the behavior of concurrent systems also note the difference between random behavior and nondeterministic behavior, in that some scheduling mechanisms work for random behavior, but not for indeterminate behavior. The main difference is that random behavior guarantees some long term characteristics that indeterminate behavior does not.

This distinction between predictability of individual sample values and predictability of long term behavior is the basis of our generalization. Our main thesis is that randomness has too much structure to describe all kinds of uncertainty.

We do not take probability as a measure of belief, and we make no philosophical justifications for using probabilities. There is nothing special about our generalization to sets, but we do require that the situations we consider can be described by a set of all possibilities. If you don’t believe probability, then you won’t believe this, and if you do, then you might not believe this anyway.

The only unusual notion we employ is for functions and relations, for which we write application on the right:

- \( A f \) is the result of applying function \( f \) to set \( A \).
- \( s^{-1} \) is the inverse relation of relation \( s \).

What we show in this paper is that there is a natural interpretation of the Dempster-Shafer belief functions [14] that not only provides us with a combined model of ignorance and probability, but also allows us to construct a (large) formal space for the non-deterministic distributions, in which the usual probability distributions become the most specific elements, that is, the ones at the top of a naturally defined partial ordering. Moreover, this partial ordering is related to the orderings in denotational semantics [15], which is the premier model of information construction in Computer Science.

2.1 Bayes’ Method for Ignorance

The Bayesian probability distribution most commonly used to model complete ignorance is the uniform probability distribution \( m \) over \( X \).

\[
m(A) = \begin{cases} 
1 / |X| & \text{for } |A| = 1, \\
0 & \text{otherwise,}
\end{cases}
\]

which is also the maximum entropy probability distribution, i.e., the probability distribution with the maximum sample value uncertainty or equivalently, the least information [1] [4].

We reject this choice as imposing structure where it is not known. This assumption of a “best guess” probability distribution is too precise. Using bounds for the values of some probability distribution is too imprecise, since it doesn’t explicitly show the relationships between the placement of probabilities of different points within their bounding intervals. We need a notion somewhere between randomness and ignorance.

2.2 Nondeterministic Distributions: Basic Definitions

In this subsection, we define the basic mathematical objects of interest. A nondeterministic distribution is a generalization of an ordinary probability distribution that allows specification of probability mass for subsets of the base set, instead of just elements. This allows us to specify certain kinds of ignorance about a probability, in addition to specifying certain kinds of knowledge about it. The objects are those of Dempster-Shafer belief functions [14], which we will assume are familiar to the reader, but the interpretation is very different.

A nondeterministic distribution on a (finite) set \( X \) is a function

\[
m : \text{powerset}(X) \to [0, 1],
\]

for which the following properties hold:

1. \( m(\emptyset) = 0 \),
2. \( \sum_{A \subseteq X} m(A) = 1 \).

We call \( X \) the reference set, or the frame of reference of the distribution. We will use a nondeterministic distribution on a frame of reference to describe incompletely known situations. The requirement for applying such a description is that the frame of reference is the set of all possible sample values, as measured in some way.

The interpretation of \( m(A) \) is the total probability known to be in \( A \) and not known to be in any proper subset of \( A \). The nondeterministic distribution \( m \) is a probability distribution if it has

\[
m(A) = 0 \quad \text{for} \quad |A| \neq 1,
\]

so that the only nonzero values of the measure are singleton subsets of \( X \), i.e., elements of \( X \).

By analogy with [14] and [6], we define two other quantities related to nondeterministic distributions. Both references use \( m \) for the nondeterministic distribution, but they call it the basic probability function.

We write \( b(A) \) for the belief in a subset \( A \) of \( X \), which is the amount of probability known to be in \( A \), and call \( b \) the belief function associated with \( m \), and \( u(A) \) for the plausibility of \( A \), which is the maximum amount of probability that could be in \( A \), and call \( u \) the plausibility function associated with \( m \). Relating these interpretations to that of \( m \), we have (see equation (2.8), p. 43, of [14])

\[
b(A) = \sum_{B \subseteq A} m(B),
\]

\[
u(A) = 1 - b(X \setminus A) = \sum_{\{B \subseteq X \mid B \text{ meets } A\}} m(B).
\]

The belief function is called that in [14], and usually represented by \( Bef \) instead of our \( b \), and is called the support in [6], and usually represented by \( s \) (or \( spf \) in [12]). The belief function \( b \) has the following properties:

1. \( b(\emptyset) = 0 \), \( b(X) = 1 \),
2. \( A \subseteq B \Rightarrow b(A) \leq b(B) \).

The last property is called monotonicity. These properties by themselves do not characterize belief functions, however.

An additional property that might be considered is that \( b \) is additive,

\[
A \cap B = \emptyset \Rightarrow b(A \cup B) = b(A) + b(B),
\]

which characterizes probability density functions among all set functions satisfying the following properties (which are special cases of the ones above):

1. \( b(A) \geq 0 \) for all \( A \subseteq X \),
2. \( b(X) = 1 \).

It is easy to show that an additive set function is always derived from a point function:

\[
b(A) = \sum_{x \in A} b(x),
\]

and that the point function \( b \) will be a probability density function when it is non-negative and the values sum to 1.

The plausibility is called that in [6], and usually represented by \( p \) instead of our \( u \), and is called the upper probability in [14], and usually represented by \( P^* \). The plausibility function \( u \) has the following properties:

1. \( u(\emptyset) = 0 \), \( u(X) = 1 \),
2. \( A \subseteq B \Rightarrow u(A) \leq u(B) \).

We can also write \( m(A) \) in terms of \( b(A) \) (Theorem 2.2, p. 39 [14]),

\[
m(A) = \sum_{B \subseteq A} (-1)^{|A \setminus B|} \cdot b(B),
\]

as is easily proved by Mobius inversion in the powerset lattice (also known as the principle of inclusion and exclusion, see Chapter 2, page 8 to 18, of [7]).

We also have (Theorem 2.8, p. 45 [14])

\[
b(A) \leq u(A) \quad \text{for all } A,
\]

and equality holds for all \( A \) if and only if \( m \) is a probability distribution. What we are calling a probability distribution has an associated belief function of the kind called a Bayesian belief function in [14], as is shown in Theorem 2.9, p. 45 (since all of the probability mass is concentrated on the singletons).

The function whose value on each \( A \subseteq X \) is

\[
u(A) - b(A)
\]

(in our notation), which is called the uncertainty \( u \) in [6], should be called imprecision or something like that, but we do not use it, since we use one summary number to measure the imprecision over all subsets of \( X \), instead of dealing with the different subsets separately.

For a nondeterministic distribution \( m \), the distribution vagueness \( V(m) \) is given by

\[
V(m) = \sum_{A \subseteq X} (u(A) - b(A)) \cdot \log(1 + |A|),
\]

which is clearly always non-negative, and zero exactly when \( m \) is a probability density function. This measure is our replacement for the uncertainty \( u \) of [6].
2.3 Uncertainty and Information

In this subsection, we describe an analogy between the uncertainty in a sample value that is usually studied in information theory and an uncertainty in distribution that is relevant to this study.

For a known fair coin, we have

\[ m(\{\text{head}, \text{tail}\}) = 0, \]
\[ m(\text{head}) = m(\text{tail}) = 1/2, \]

and for an unknown coin, we have

\[ m(\{\text{head}, \text{tail}\}) = 1, \]
\[ m(\text{head}) = m(\text{tail}) = 0, \]

and we want a numerical measure of uncertainty that can distinguish between the two cases. In the former case, we have some uncertainty in the individual outcomes, but some guarantees of the asymptotic behavior. In the latter case, we have no guarantees of anything.

We choose to measure our uncertainty in the distribution as a different kind of uncertainty than that of the sample value uncertainty of the fair coin. We will take the first case to have no uncertainty of distribution, and in fact, any probability distribution will have no uncertainty of distribution.

2.3.1 Information Theory

We recall the information theory definitions of uncertainty and entropy for a probability distribution [1]. The entropy of a probability distribution \( p \) over a finite set \( X \) is

\[ H(p) = - \sum_{x \in X} p(x) \cdot \lg p(x), \]

with the convention that

\[ 0 \cdot \lg 0 = 0, \]

where \( \lg(x) \) is the base 2 logarithm. The entropy is also called uncertainty, meaning uncertainty of individual outcomes, but we will reserve the term “uncertainty” for distribution uncertainty, and use the term entropy for the usual information theory entropy.

The minimum value of \( H(p) \) is 0, which occurs whenever \( p(x) \) is zero for all but one element \( x \in X \). The maximum value of \( H(p) \) is \( \lg |X| \) which occurs when \( p \) is the uniform distribution over \( X \) [1].

These properties are here for comparison with the properties derived below for non-deterministic distributions. They show that our definitions are quite compatible with classical information theory, and even motivate some of the definitions.

2.3.2 Distribution Uncertainty

Our main measure of uncertainty is the distribution uncertainty \( U(m) \) of a nondeterministic distribution \( m \), which is defined by

\[ U(m) = \sum_{A \subseteq X} m(A) \cdot \lg |A|, \]

which is clearly non-negative. This measure has value zero for any probability distribution, since in that case \( m(A) \) is nonzero only for \( |A| = 1 \), and maximum value \( \lg(|X|) \) for the indeterminate distribution

\[ m_X(X) = 1, \]
\[ m_X(A) = 0 \text{ for } A \neq X, \]

as can be easily verified.

**Lemma.** The maximum value of \( U(m) \) is \( \lg(|X|) \).

Before we prove this lemma, we describe a construction that will be used several times in subsequent proofs. Take any pair of sets

\[ A \subseteq B \subseteq X, \]

for which \( m(B) > 0 \), and any positive \( d \leq m(B) \), and define a new function \( m' \) by

\[ m'(A) = m(A) + d, \]
\[ m'(B) = m(B) - d, \]
\[ m'(C) = m(C) \text{ for all other } C, \]

and compare \( U(m') \) to \( U(m) \),

\[ U(m') = \sum_{C \subseteq X} m'(C) \cdot \lg |C| \]
\[ = \sum_{C \subseteq X} m(C) \cdot \lg |C| \]
\[ + (m(A) + d) \cdot \lg |A| \]
\[ + (m(B) - d) \cdot \lg |B| \]
\[ = \sum_{C \subseteq X} m(C) \cdot \lg |C| \]
\[ + d \cdot \lg |A| \]
\[ - d \cdot \lg |B| \]
\[ = U(m) + d \cdot (|A| - |B|) \]
\[ < U(m), \]

since \( |A| \) is smaller than \( |B| \).

**Proof of Lemma.** Therefore, the more restricted a distribution is, in the general sense of having more of its probability measure constrained to smaller sets, the less its uncertainty is. It is easily seen from this
comparison that the uncertainty is largest for the indeterminate distribution. We will define a precise ordering of specificity later, and show that it refines this numerical notion of uncertainty.

We define the information contained in a nondeterministic distribution over \( X \) to be

\[
I(m) = \lg |X| - U(m),
\]

\[
= - \sum_{A \subseteq X} m(A) \star \lg \frac{|A|}{|X|},
\]

so that the indeterminate distribution contains zero information, and any probability distribution contains the maximum amount of information, which is \( \lg |X| \).

2.4 Specificity

In this subsection, we define a notion of specificity of nondeterministic distribution, and relate it to the idea of specificity in denotational semantics [15]. We will define a partial ordering between nondeterministic distributions, and show that the set of all nondeterministic distributions is a chain complete partially ordered set, which is one in which every non-decreasing chain has a least upper bound.

2.4.1 The Partial Ordering

For two nondeterministic distributions \( m_1 \) and \( m_2 \), we write

\[
m_1 \leq m_2,
\]

read \( m_1 \) is less specific than \( m_2 \), if and only if for all \( A \subseteq X \),

\[
b_1(A) \leq b_2(A),
\]

which is clearly a partial ordering between nondeterministic distributions, where we have written \( b_1 \) for the belief function derived from \( m_1 \) and similarly \( b_2 \) for \( m_2 \). The partial order condition is equivalent to requiring

\[
u_1(A) \geq u_2(A),
\]

for all \( A \subseteq X \), where we have written \( u_1 \) for the plausibility function derived from \( m_1 \), and similarly \( u_2 \) for \( m_2 \). We define more specific by reversing the inequalities, and we may write the inequalities between distributions, belief functions, or plausibility functions. It should be noted in this context that the inequalities for the plausibility functions are reversed.

It is also easy to show, using the construction of the lemma in Section 2.3.2, that if \( m_1 \leq m_2 \), then

\[
b_1 \leq b_2 \leq u_2 \leq u_1,
\]

so that the distribution vagueness has

\[
V(m_2) \leq V(m_1).
\]

We write \( P(X) \) for the set of all nondeterministic distributions, under the above partial ordering.

2.4.2 Minimum and Maximal Elements

In this subsection, we will determine the extreme elements of the partially ordered set \( P(X) \).

**Lemma.** There is only one minimal element of \( P(X) \), so we can call it the minimum element.

**Proof.** It is easily seen that since every belief function \( b \) has

\[
b(X) = 1,
\]

\[
b(A) \geq 0 \text{ for } A \subseteq X,
\]

we have that the indeterminate distribution \( b_2 \) is the least specific, since it has

\[
b_2(A) = 0 \text{ for } A \subset X,
\]

\[
b_2(X) = 1.
\]

Therefore, the indeterminate distribution \( b_2 \) is the unique minimal element of \( P(X) \).

Now we find the maximal (most specific) elements of \( P(X) \). It turns out that there are many of them. Some of these results are proved in [13], in Theorem 4.3, p. 40, though the intent there is limited to showing a relationship between probabilities and Dempster-Shafer belief functions. We think that the notion of trying to characterize a belief function by the collection of all probability distribution functions that are consistent with it is a fruitful area of inquiry, because we will have a lattice with some geometric properties.

**Lemma.** The probability distributions over \( X \) are maximal elements of \( P(X) \).

**Proof.** Let \( p_1 \) be a probability distribution, \( m_2 \) a nondeterministic distribution, and suppose that

\[
p_1 \leq m_2
\]

so that for all \( A \subseteq X \),

\[
b_1(A) \leq b_2(A).
\]

If we consider these inequalities only for singleton sets \( A \), and note that \( b_2(\emptyset) \) is zero, we have that for all \( x \in X \),

\[
p_1(x) \leq m_2(\{x\}).
\]

Now since the sum over all \( x \in X \) of all the terms on the left is one, and the sum of all the terms on
the right is at most one, the inequalities collapse, and
\( m_2 \) and \( p_1 \) are the same for singletons. Since the sum
over all subsets of \( A \) of \( m_2(A) \) can only be one, the
other terms must all be zero, so we have \( m_2 \) is also a
probability distribution, and therefore that \( m_2 \) and \( p_1 \)
are the same.

**Lemma.** The probability distributions over \( X \) are
all of the maximal elements of \( P(X) \).

**Proof.** It suffices to show that a nondeterministic
distribution \( m_1 \) that is not a probability distribution
is not maximal. So let \( m_1(A) \) be positive for some set
\( A \subseteq X \) with \( |A| > 1 \), and define the nondeterministic
distribution \( m_2 \) by

\[
\begin{align*}
m_2(A) &= 0, \\
m_2(B) &= m_1(A) + m_1(B),
\end{align*}
\]

for any fixed but arbitrary nonempty subset \( B \subseteq A \),
and

\[ m_2(C) = m_1(C) \]

for any subset \( C \subseteq X \) that is not either \( A \) or \( B \).

Then we have that

\[ b_2(C) = b_1(C) \]

for any subset \( C \subseteq X \) that either contains \( A \), since the
\( m_1(A) \) term appears in \( m_2(B) \), or does not contain \( B \),
since the different terms do not appear in the sum.

It remains to consider subsets \( C \subseteq X \) that do not
contain \( A \) and do contain \( B \). We have

\[ b_2(C) = b_1(C) + m_1(A), \]

so that

\[ b_2 \geq b_1, \]

and \( m_1 \) is not maximal.

It should be noted that the proofs of the lemmas in
Section 2.3.2, Section 2.4.1, and this section all use
the same construction.

**2.4.3 Uncertainty and the Partial Ordering**

We now relate the ordering in \( P(X) \) to the distribution
uncertainty function defined earlier.

**Lemma.** For two nondeterministic distributions \( m_1 \)
and \( m_2 \), if

\[ m_1 \leq m_2, \]

then the distribution uncertainties have

\[ U(m_1) \geq U(m_2), \]

so that the less specific distributions have more uncertain-
ity.

**Proof.** The best proof we have so far for this lemma
is extremely ugly, so it is omitted.

**2.4.4 Chains of Distributions**

Now we consider a non-decreasing chain

\[ b_0 \leq b_1 \leq b_2 \leq \ldots \]

of more and more specific belief functions in \( P(X) \), and
we show that there is a least upper bound in \( P(X) \). It
will follow that every element of \( P(X) \) is bounded by
(less specific than) a maximal element (some probability
distribution).

Note that for each \( A \subseteq X \), the sequence

\[ (b_i(A) \mid i \geq 0) \]

is a non-decreasing sequence of non-negative real numbers, all of which are at most 1. Therefore,

\[ b(A) = \lim_{i \to \infty} b_i(A) \]

exists and is at most 1.

Since each \( m_i(A) \) is a finite sum of terms of the form
\( b_i(B) \), we also have

\[ m(A) = \lim_{i \to \infty} m_i(A) \]

exists and has the required relationship to the terms
\( b(B) \). Moreover, since the limits exist, it is easy to
show that \( m \) is also a nondeterministic distribution.

**Theorem.** The set \( P(X) \) of all nondeterministic
distributions on \( X \) is a chain complete partially ordered
set.

This kind of partially ordered set is used in deno-
tational semantics [15] to describe various amounts of
missing information in the specification of computa-
ions.

**2.4.5 Consistency and the Partial Ordering**

There is an obvious notion of consistency between
a nondeterministic distribution and a probability distri-
bution, based on the notion that the former could
be changed by new information into the latter. We show
that this relationship is the same as the partial
ordering in \( P(X) \).

A nondeterministic distribution \( m_1 \) is *consistent*
with the probability distribution \( p_2 \) if and only if the
amount of probability committed to a subset \( A \subseteq X \) by
\( p_2 \) does not exceed the total amount of belief committed
to sets intersecting \( A \) by \( m_1 \), since their measure
could be moved to the common elements.

We can easily restate the condition using the related
functions. For all \( A \subseteq X \),

\[
\sum_{x \in A} p_2(x) \leq \sum_{B \subseteq X \mid B \text{ meets } A} m_1(B),
\]

so that the less specific distributions have more uncertain-
ity.

**Proof.** The best proof we have so far for this lemma
is extremely ugly, so it is omitted.
and since we have

$$u_2(A) = \sum_{x \in A} p_2(x),$$

we have $m_1$ is consistent with $p_2$ if and only if $u_1 \geq u_2$ if and only if $b_1 \leq b_2$. Therefore, a nondeterministic distribution $m_1$ is consistent with the probability distribution $p_2$ if and only if $m_1 \leq p_2$ in $P(X)$. Since this notion of consistency is a special case of the partial ordering by specificity, we can also write $m_1$ and $m_2$ are consistent if and only if either

$$m_1 \leq m_2 \text{ or } m_1 \geq m_2.$$

This relation is reflexive and symmetric, but not transitive.

2.5 Best Guess

For a given nondeterministic distribution, it is often convenient to have a representative probability distribution to work with, despite the fact that we rejected the view that there is a probability distribution that completely represents our state of knowledge. We will allow that there is a representative probability distribution for the state of our knowledge, but that the state of our ignorance can only be measured by something like a nondeterministic distribution.

We define the best guess probability distribution $p$ for a given nondeterministic distribution $m$ by

$$p(x) = \sum_{\{A \subseteq X\mid x \in A\}} m(A) / |A|,$$

which spreads out each $m(A)$ over all elements of $A$ equally, conforming to the Bayesian response to ignorance.

**Lemma.** The best guess probability distribution has

$$\sum_{x \in A} p(x) = \sum_{\{B \subseteq X\mid |B| \neq 0\}} \left| \frac{A \cap B}{|B|} \right| m(B)$$

for all $A \subseteq X$.

**Proof.** Easy.

**Lemma.** For the indeterminate distribution $m_1$, the best guess is the uniform probability distribution, and for any probability distribution $p$, the best guess is just $p$.

**Proof.** Easy.

**Lemma.** The best guess probability distribution is consistent with the nondeterministic distribution from which it is computed.

**Proof.** Easy from the first lemma of this subsection, since

$$\frac{|A \cap B|}{|B|} \leq 1$$

for all $A, B \subseteq X$ with $B$ nonempty.

3 Extensions of the Basic Theory

In this section, we make a short foray into two extensions of the basic theory, first examining some ways to change a frame of reference, then considering some ways to treat inconsistency. The extensions broaden the scope of applicability of our model.

3.1 Changing the Frame of Reference

In this subsection, we start with some ways to change the frame of reference for a nondeterministic distribution. This operation is an analogue of gaining new information about the framing of a situation. The reason that it is important to be able to compute changes of reference is that many errors in probabilistic arguments occur because either the framing does not properly match the problem, or information defined in one frame is combined with information defined in some other frame. Having some explicit methods for computation of frame changes should help reduce the occurrence of these errors.

3.1.1 Simplification

Suppose we have a nondeterministic distribution $m$ on $X$, describing a certain situation, and suppose that we discover new information that causes us to expand the set $X$ into a larger set $X'$ by splitting up the elements of $X$. This separation of each element of $X$ into one or more elements of $X'$ can be described by a function

$$s : X' \rightarrow X,$$

called a simplification function, assumed to be surjective. Then each element $x \in X$ expands to the set $x.s^{-1} \subseteq X'$. Simplification has the following properties:

1. $x.s^{-1}$ is nonempty,

2. $x, y$ different implies $x, s^{-1}, y, s^{-1}$ disjoint.

Under these conditions, we can transfer $m$ to a nondeterministic distribution $m'$ on the larger set $X'$ by
defining
\[ m'(A') = m(A) \text{ if } A' = A.s^{-1}, \]
\[ m'(A') = 0 \text{ otherwise,} \]
so that \( m'(A') \) is nonzero only if \( A' \) is the entire preimage of \( A'.s \), and if \( m(A'.s) \) is nonzero also.

We can also transfer in the other direction. If \( s \) is a simplification and \( m' \) is a nondeterministic distribution on \( X' \), then we define a nondeterministic distribution \( m \) on \( X \) by
\[
m(A) = \sum_{\{A' \subseteq X' : A'.s = A\}} m'(A').
\]
Since each \( m'(A') \) term occurs in exactly one sum, the resulting function \( m \) is a nondeterministic distribution on \( X \).

**Lemma.** If \( m' \) was originally constructed using \( s \) from a nondeterministic distribution \( m'' \) on \( X \), then \( m = m'' \).

**Proof.** Easy.

Note that for any function \( s \), and any finite set \( A' \) in the domain of definition of \( s \), we have
\[ |A'.s| \leq |A'| \]
with equality iff \( s \) is injective on \( A' \).

**Lemma.** \( U(m) \leq U(m') \). The lemma can be paraphrased as saying that simplification makes a distribution more precise (not necessarily more accurate, though).

**Proof.** We have
\[
U(m) = \sum_{A \subseteq X} m(A) \cdot \log |A| = \sum_{A \subseteq X} \sum_{A'.s = A} m'(A') \cdot \log |A| = \sum_{A' \subseteq X'} \sum_{A'.s = A} m'(A') \cdot \log |A'.s| \leq \sum_{A' \subseteq X'} m'(A') \cdot \log |A'| = U(m').
\]

There is no obvious relationship between the new best guess and the old one, since the assumptions that produced them are different. It is like suddenly finding that the coin you thought had two sides actually has two different “heads". If you don’t know that the two choices are distinguishable by their behavior (not necessarily by your observations), then you must assume otherwise, giving two choices for results, and if you do know that the two choices are distinguishable (whether or not you can distinguish them), then you must allow all three choices for results.

The best guess probability distribution in the first case is \( 1/2 \) for heads and \( 1/2 \) for tails, but in the second case it is \( 2/3 \) for heads and \( 1/3 \) for tails. This difference is part of the reason we prefer to avoid the best guess.

### 3.1.2 Simple Refinement

Simplification is a kind of disjoint refinement, in that different elements of a smaller set \( X \) become disjoint subsets of the refined set \( X' \). A more general kind of refinement takes a map
\[
r : X \rightarrow \text{powerset}(X'),
\]
and calls it a refinement function when it has the following properties:

1. \( r(x) \neq \emptyset \) for all \( x \in X \).
2. for different \( x, y \in X \), \( r(x) \) is not a subset of \( r(y) \).

The second condition is related to the coherence of our interpretation of a nondeterministic distribution, since it assumes that further information will not cause one previously distinguishable alternative to disappear entirely within another. Note that the first condition is implied by the second condition when \( X \) has at least two elements. The second condition implies that \( r \) is injective. This kind of refinement is clearly a generalization of simplification, since the disjoint refinement derived from a simplification function \( s \) is obtained by taking \( r \) above to be \( s^{-1} \).

We extend any such function \( r \) to a function
\[
r : \text{powerset}(X) \rightarrow \text{powerset}(X')
\]
in the usual way, by defining
\[
r(A) = \bigcup_{x \in A} r(x).
\]
The extended function is monotonic:
\[
A \subseteq B \Rightarrow r(A) \subseteq r(B).
\]

For any nondeterministic distribution \( m \) on \( X \), we can use the extended function \( r \) to define a nondeterministic distribution \( m' \) on \( X' \) by
\[
m'(A') = \sum_{\{A \subseteq X : A'.s = A\}} m(A),
\]
so \( m'(A') \) is nonzero if and only if \( A' \) is the image of some \( A \) with \( m(A) \) nonzero. Since each \( m(A) \) term appears in exactly one \( m'(A') \) sum, the resulting function \( m \) is a nondeterministic distribution.
We can also consider the reverse operation. For $m'$ on $X$, define $m$ on $X$ by
\[ m(A) = \sum_{\{A' \subseteq X' \mid A' \sim A, r\}} m'(A'). \]
In order for this process to map non-deterministic distributions on $X'$ into non-deterministic distributions on $X$, $r$ must be injective, at least on sets $A$ with $m'(A, r) > 0$. Otherwise, more than one $m(A)$ gets a contribution from $m'(A')$ (so the sum of all $m(A)$ will generally be too large). Moreover, $r$ must be surjective on sets $A'$ with $m'(A') > 0$, since otherwise, no $m(A)$ will get a contribution from $m'(A')$ (so the sum of all $m(A)$ will generally be too small).

In this case also, the best guess has no obvious transformation corresponding to the refinement function.

3.2 Inconsistency

In this subsection, we explore the utility of the idea of an inconsistent distribution. We consider several different forms of inconsistency, based on relaxing the conditions on different functions.

3.2.1 Distribution Inconsistency

We first remove the condition on nondeterministic distributions that $m(\emptyset)$ is zero, and call $m(\emptyset)$ the level of inconsistency of $m$. The general approach will be to consider an inconsistent distribution as grounds for changing the frame of reference.

We consider extensions to the frame of reference as a mechanism remove an inconsistency. We take
\[ m'(A) = m(A) \]
for nonempty $A \subseteq X$, and
\[ m'(x) = m(\emptyset) \]
for a new element $x \notin X$.

The new element $x$ was introduced as a kind of indication that something was wrong with the original frame selection. We do not yet have any constructions of inconsistent nondeterministic distributions that can be fixed using this kind of extension.

3.2.2 Normalization to Remove Inconsistency

An alternative procedure is to "throw out" the inconsistent part of the probability. In this case, we normalize $m$ to
\[ m'(A) = m(A)/(1 - m(\emptyset)) \]
for nonempty $A \subseteq X$, which usually gives us an ordinary nondeterministic distribution. Then the only trouble is the completely inconsistent distribution that has value 1 at $\emptyset$, 0 otherwise. If the observations have this distribution, nothing interesting can happen anyway, so we ignore it.

Note that the rules of combination for evidence [14] can produce this kind of inconsistent function, and that normalization is the preferred method of removing the inconsistency. Similarly, conditional belief functions are defined as though they are normalized versions of a possibly inconsistent difference [13].

3.2.3 Belief Inconsistency

Another kind of inconsistency occurs with belief functions. For this investigation, we define a generalized belief function on $X$ to be a function
\[ b : \text{powerset}(X) \to [0, 1], \]
for which the following conditions hold, as they do for ordinary belief functions:
1. $b(\emptyset) = 0$, $b(X) = 1$,
2. $A \subseteq B \Rightarrow b(A) \leq b(B)$.

We call the last condition monotonicity. If the equation that normally relates the belief function to a nondeterministic distribution function is used, we can define a function
\[ m : \text{powerset}(X) \to \text{Real numbers} \]
by
\[ m(A) = \sum_{B \subseteq A} (-1)^{|A| - |B|} \cdot b(B), \]
for all $A \subseteq X$. The inverse equation, relating $b$ to a sum of values of $m$, remains true for these functions. Since nothing in the definition of the generalized belief function prevents it, some of the values $m(A)$ may be negative (the sum of all of them is $b(X)$, which is still 1). In such cases, we call the generalized belief function inconsistent, and consider changes in the frame of reference to make the transformed belief function consistent.

There may be simple arithmetic adjustments that fix the values of $m$, but we are looking for some change in the frame that is more natural. In particular, we would like some kind of refinement or simplification to fix the problem, but we don’t have a good answer for now. The way we interpret a subset $A$ with $m(A) < 0$ is that $m$ is claiming that more of the probability is
in the subsets of $A$ than is possible. What that ought to mean is that some of those subsets are the same, so that they are counted twice, and therefore that some of the elements of $A$ should be identified, or that the subsets should be mapped in a more complicated way.

4 Conclusions

This paper is almost entirely theoretical, providing the basic foundations of a new uncertainty model. It combines ignorance in a simple but ad hoc manner with the usual probability model. We have connected this model to information theory and to denotational semantics, which is one of the main models of information in Computer Science.

We are currently applying the model to some difficult data analysis problems [9], and using the same method to combine ignorance with other uncertainty models. We do not believe that a single comprehensive model is possible in principle [2], so we are exploring the use of explicit knowledge-based integration methods to allow the separate models to cooperate [10].

References


