2nd International Symposium on Imprecise Probabilities and Their Applications, Ithaca, New York, 2001

# Confidence as Higher-Order Uncertainty

**Pei Wang**
Center for Research on Concepts and Cognition, Indiana University
pwang@cogsci.indiana.edu
http://www.cogsci.indiana.edu/farg/peiwang/

## Abstract

With conflicting evidence, a reasoning system derives uncertain conclusions. If the system is open to new evidence, it faces additionally a higher-order uncertainty, because the first-order uncertainty evaluations are uncertain themselves — they can be changed by future evidence. A new measurement, *confidence*, is introduced for this higher-order uncertainty. It is defined in terms of the amount of available evidence, and interpreted and processed as the relative stability of the first-order uncertainty evaluation. Its relation with other approaches of "reasoning with uncertainty" is also discussed.

**Keywords.** confidence, evidence, frequency interval, revision, inference, deduction, induction, abduction.

## 1 Introduction

In this paper, we discuss the representation and processing of uncertainty in an adaptive reasoning system, whose knowledge and resources are insufficient with respect to the questions to be answered.

NARS (for Non-Axiomatic Reasoning System, see [26] for details) is a reasoning system that accepts knowledge and questions from its user in a formal language, and answers the questions according to available knowledge. No restriction is imposed on the contents of the knowledge and questions the system may encounter, as long as they are expressible in the formal language. Therefore, new knowledge may come from time to time, and may conflict with previous knowledge. Also, questions may go beyond the system's current knowledge scope.

We want the system to be adaptive, that is, to behave according to its experience (available knowledge/evidence). In such a situation, the system's answers are usually uncertain, since the input knowledge is not necessarily conflict-free, and the system needs to make plausible inferences when the available knowledge is insufficient to answer a question with absolute certainty. As a result, for a given question, NARS usually cannot find a unique "correct" or "optimal" answer, but only a "reasonable" answer that is best supported by its experience, and can be found or constructed under the current time-space constraint.

In the following sections, we explain why in NARS the representation of uncertainty needs *two* numbers. A new measurement, *confidence*, is introduced, and some operations on this measurement are discussed. Finally, this approach is compared with other probability-based measurements of higher-order uncertainty.

Since a comprehensive introduction to NARS is far beyond the capacity of this paper, we will focus on the confidence issue, and for the other aspects of the system (such as knowledge representation language, semantics, inference rules, truth value functions, memory management, inference control, and so on), we only mention the directly relevant parts. Papers and an on-line demonstration of NARS are available at the author's web-page.

## 2 Evidence and Confidence

In NARS, the basic form of knowledge is an *inheritance relation* between two terms, "$S \subset P$". Intuitively, it indicates that $S$ is a *specialization* of $P$, and $P$ is a *generalization* of $S$. It roughly corresponds to "$S$ is a kind of $P$" in English.

In its ideal form, Inheritance is reflexive and transitive. In NARS, the *extension* and *intension* of a term $T$ are defined as sets of terms: $E_T = \{x \mid x \subset T\}$ and $I_T = \{x \mid T \subset x\}$. Intuitively, they include all known specialization (instances) and generalizations (properties) of $T$, respectively.

It can be proven that $(S \subset P) \Longleftrightarrow (E_S \subseteq E_P) \Longleftrightarrow (I_P \subseteq I_S)$, where the first relation is an Inheritance relation between two terms, while the last two are subset relations between two sets (extensions and intensions of terms). Therefore, $S \subset P$ indicates that $S$ inherits the intension of $P$, and $P$ inherits the extension of $S$.

Given the assumption of insufficient knowledge and the requirement of being adaptive, the uncertainty of $S \subset P$ is defined according to available evidence in the system. For a statement $S \subset P$ and a term $M$, if $M$ is in the extensions of both $S$ and $P$, it is positive evidence for the statement; if it is in the extensions of $S$ but not the extension of $P$, it is negative evidence. Symmetrically, if $M$ is in the intensions of both $P$ and $S$, it is positive evidence for the statement; if it is in the intension of $P$ but not the intension of $S$, it is negative evidence. Therefore, the amount of positive evidence is $w^+ = |E_S \cap E_P| + |I_P \cap I_S|$, the amount of negative evidence is $w^- = |E_S - E_P| + |I_P - I_S|$, and the amount of all evidence is $w = w^+ + w^- = |E_S| + |I_P|$.

For example, an observed black crow is a piece of positive evidence for "Crow is a kind of black thing" ($w = w^+ = 1$), and an observed non-black crow is a piece of negative evidence for it ($w = w^- = 1$). Here we assume the observations have no uncertainty.

To measure the amount (or weight) of evidence is not a new idea at all [10, 15]. For instance, Keynes said that "As the relevant evidence at our disposal increases, the magnitude of the probability of the argument may either decrease or increase, according as the new knowledge strengthens the unfavorable or the favorable evidence; but *something* seems to have increased in either case, — we have a more substantial basis upon which to rest our conclusion." [10]

Though we do not always make a judgment by directly counting pieces of evidence, the concept of *amount of evidence* can be used as an idealized meter-stick by which uncertainty is measured — I can say that my belief on a sentence is "as strong as I have tested the sentence $w$ times, and the tests succeeded $w^+$ times, but failed $w^-$ times", even though I did not really test the sentence in this way. For how to apply this measurement to a formal language, see [25, 26].

Because all the operations in the system are based on available evidence, $w^+$ and $w^-$ contain all the information about the uncertainty of the sentence. However, when represented in this way, the information is inconvenient for certain purposes. When comparing competing options and deriving new conclusions, we usually prefer *relative measurements* to *absolute measurements*.

The most often used relative measurement for uncertainty is the *frequency*, or *proportion*, of positive evidence among all available evidence. In the following, let us define the "frequency" of a sentence as $f = w^+/w$. If the system has observed 100 crows, and 95 of them are black, but the remaining 5 are not, the system sets $f = 0.95$ for "Crow is a kind of black thing".

Although $f$ is a natural and useful measurement, it is not enough for our current purpose. Intuitively, we have the feeling that the uncertainty evaluation $f = 0.95$ is uncertain itself. For a simple example, let us consider the following two situations: (1) the system only knows one crow, and it is black, and (2) the system knows 10000 crows, and all of them are black. Though in both situations we have $f = 1$, the first case is obviously "more uncertain" than the second. Because here the uncertainty is about the sentence "The frequency for crows to be black is 1", we are facing a *higher-order* uncertainty, which is the uncertainty of an evaluation about uncertainty.

As mentioned at the beginning of the paper, in NARS the uncertainty in a sentence appears as the result of insufficient knowledge. Specially, the first-order uncertainty, measured by frequency, is caused by *known* negative evidence, and the higher-order uncertainty is caused by *potential* negative evidence.

As discussed previously, the most simple and natural measurement of the higher-order uncertainty is the amount of evidence, $w$. However, we have reasons to introduce a *relative* measurement, whose advantages will be apparent later.

Intuitively, we are looking for a function of $w$, call it $c$ for *confidence*, that satisfies the following conditions:

1. Confidence $c$ is a continuous and monotonically increasing function of $w$. (More evidence, higher confidence.)

2. When $w = 0$, $c = 0$. (Without any evidence, confidence is minimum.)

3. When $w$ goes to infinity, $c$ converges to 1. (With infinite evidence, confidence is maximum.)

There are infinite functions satisfying the above requirements, therefore we need more intuition to pick up a specific one.

Many functions with value range [0, 1] can be naturally interpreted as a *ratio*. Following this path, we might want to define $c$ as the ratio of "the amount of evidence the system has known" to "the amount of evidence the system will know". Obviously, the first item is $w$, but for a system that is always open to new evidence, the second item is infinity, therefore the ratio is always 0. When compared with an infinite "future," the difference among the various finite "past" cannot be perceived. For an adaptive system, though past experience is never sufficient to predict future situations, the amount of evidence does matter for the system's decision, and the behaviors based on more evidence should be preferred.

Because confidence is supposed to be a *relative* measurement defined on finite evidence, a natural idea is to compare the amount of current evidence with the amount of evidence *the system will know in the near future*. By "near future," we mean "until the coming of a constant amount of new evidence".

Now we get the definition of confidence that we want to introduce in this paper: $c = w/(w + k)$, where $k$ is a positive constant indicating the "near future". Defined in this way, the frequency and confidence of a sentence are independent to each other, in the sense that, from the value of one, the other's value cannot be determined, or even estimated or bounded (except the trivial case where $c = 0$ indicates that $f$ has an undefined value).

Obviously, this function satisfies the three requirements. For a specific system, $k$ should remain unchanged to make the system's behaviors consistent, but different systems can have different values for $k$. In this paper, the default value of $k$ is 1 (and we will discuss the choice of $k$ later). Under such a definition, confidence indicates the ratio of the current amount of evidence (for the piece of knowledge under consideration) to the amount of evidence the system will have after it gets new evidence with a unit amount (or weight). The more the system already knows, the less the new evidence will contribute (relatively), therefore the more confident, or the less ignorant, the system is, with respect to the given relation. When $w = 1$, $c = 0.5$, and the new evidence will double the amount of available evidence; When $w = 999$, $c = 0.999$, and the new evidence will have little effect on the system's belief.

For empirical knowledge, $c$ can never reach 1, so the knowledge is always (more or less) revisible. In NARS, $c = 1$ is reserved for *analytical knowledge*, such as mathematical knowledge. This kind of knowledge is not a direct summary of experience, but a convention that is not directly revisible by evidence (we will return to this issue later).

We can interpret confidence in another way. As defined previously, the current frequency of positive evidence is $f = w^+/w$. After getting a piece of new evidence with weight $k$, where will the new $f$ be? Obviously, if the new evidence is completely negative, $f$ will be $w^+/(w + k)$; if the new evidence is completely positive, $f$ will be $(w^+ + k)/(w + k)$. Therefore, no matter what content the new evidence has, the frequency will stay in the interval $[w^+/(w + k), (w^+ + k)/(w + k)]$ in the near future. Let us call the lower bound and the upper bound of the interval "lower frequency" and "upper frequency", respectively. The *width* of the interval, $k/(w + k)$, provides a measurement for the ignorance (or susceptibility) of the system on the statement, which is a higher-order uncertainty, and its complement (to 1), $w/(w+k)$, provides a measurement for the confidence (or stability) of the judgment.

Now we have three functionally identical ways to represent the uncertainty of a statement: by two of the three amounts of evidence ($w^+$, $w^-$, and $w$), by the two ratios (frequency and confidence), or by the lower–upper frequency interval. From the above definitions, it is not difficult to get the one-to-one mappings among the three representations [25, 26]. No matter which form is used, we need *two* numbers to represent the uncertainty of a statement.

Is this kind of information available to the system? Even Bayesian network and fuzzy logic, which require the users to assign *a single number* to each statement, have difficulty in getting the numbers. How can we now expect users to provide *a pair of numbers* for each statement? To us, the hardness of value assignments comes mainly from the unclear interpretation about *what are measured by these values*. Our approach attempts to be more user friendly by unifying different uncertainty representations. This clarifies the assignment process for the users. They can even mix different forms of uncertainty, in terms of amount of evidence, frequency, confidence, ignorance, frequency interval, and so on, in the knowledge they provide.

Though $c$ is defined as a function of $w$, which is intuitively understood as "amount of available evidence," the system does not simply count the number of pieces of evidence, and treat each of them as equally-weighted. Instead, it interprets evidence as equivalent to ("as strong as") that from such a simple counting. If a statement has a confidence $c = 0.9$, which corresponds to $w = 9$ when $k = 1$, it does not mean that the system really has found 9 pieces of (ideal) evidence, but that the system believes the statement to such an extent, *as if* it has found 9 pieces of ideal evidence for the statement [26].

## 3 Confidence-related Operations

In the following, we show how confidence is processed in the major inference operations of NARS.

### 3.1 Revision

In NARS, *revision* indicates the process by which evidence from different sources is combined. For example, assuming the system's previous uncertainty for "Crows are black" is $<9/10,\ 10/11>$ (we know that it corresponds to "Ten crows observed, and nine of them are black" when $k = 1$), now a piece of new knowledge comes, which is "Crows are black $< 3/4, 4/5 >$" (so it corresponds to "Four crows are observed, and three of them are black"). If the system can determine that no evidence is repeatedly counted in the two sources (see [25, 26] for how this is defined and checked), then the uncertainty of the revised judgment should be $<6/7,\ 14/15>$ (corresponding to "Fourteen crows observed, and twelve of them are black").

Formally, revision is the rule that merges $< f_1,\ c_1 >$ and $< f_2,\ c_2 >$ into $< f,\ c >$ for the same $S \subset P$. From the conversion that the amount of evidence is additive during revision and the definition of frequency and confidence in terms of evidence, we get the following uncertainty function for the revision rule:

$$
\begin{aligned}
f &= \frac{f_1 c_1(1-c_2)+f_2 c_2(1-c_1)}{c_1(1-c_2)+c_2(1-c_1)} \\
c &= \frac{c_1(1-c_2)+c_2(1-c_1)}{c_1(1-c_2)+c_2(1-c_1)+(1-c_1)(1-c_2)}
\end{aligned}
$$

where $< f_1,\ c_1 >$ and $< f_2,\ c_2 >$ are the uncertainty of the two premises, and $< f,\ c >$ is the uncertainty of the conclusion.

This function has the following properties:

1. The order of the premises does not matter.
2. As a weighted average of $f_1$ and $f_2$, $f$ is usually a "compromise" of them, and is closer to the one that is supported by more evidence.
3. $c$ is never smaller than either $c_1$ or $c_2$, that is, the conclusion is supported by no less evidence than either premise.
4. If $c_1 = 0$, then $f = f_2$ and $c = c_2$, that is, a judgment supported by null evidence cannot revise another judgment.
5. If $c_1 = 1$ and $c_2 < 1$, then $f = f_1$ and $c = c_1$, that is, a definition (supported by complete evidence) cannot be modified by empirical evidence.
6. If $c_1 = c_2 = 1$ but $f_1 \neq f_2$, then $f$ and $c$ are undefined, that is, there are two conflicting definitions in the system, from which nothing can be derived.

This definition is compatible with our intuition about evidence and revision — revision is nothing but to reevaluate the uncertainty of a statement by taking new evidence into account. Revision is not updating, where old evidence is thrown away. A high $w$ means that the system already has much evidence for the statement, therefore its confidence is high and its ignorance is low (on this issue). It follows that the statement is relatively insensitive to new evidence. All these properties are independent to the decisions on how $w$ is divided into $w^+$ and $w^-$, as well as to how they are actually measured (so these decisions may change from situation to situation without invalidating the revision rule).

It needs to be clarified that here "revision" refers to the operation by which the system summarize two (maybe conflicting) beliefs. In this operation the conclusion always has a higher confidence. However, generally speaking, in NARS it is possible for the system to loss its confidence on a belief. This can be caused by the "forgetting" or "explaining away" of previously available evidence. This issue is related to the memory management and control mechanism of the system [26], and thus is beyond the scope of this paper.

### 3.2 Choice

What will happen if the evidence of the two premises are "correlated", that is, some evidence are used by both of them? Ideally, we would like to merge the evidence without repeatedly counting the overlapping part. However, with insufficient resources (which is assumed in NARS), it is simply impossible to distinguish the contribution of each piece of evidence to the uncertainty of the judgment. Therefore, when NARS recognizes that two premises are based on correlated evidence (for how this can be done, see [26]), it chooses the premise with a higher *confidence*, because it is supported by more evidence. This is exactly what we expect from an adaptive system whose behaviors are based on its experience.

Another type of choice happens when the competing statements have different contents. For example, the system needs to decide which one of two statements $S_1 \subset P < f_1,\ c_1 >$ and $S_2 \subset P < f_2,\ c_2 >$ is more likely to be confirmed in the next time they are tested (i.e., which of $S_1$ and $S_2$ is a better candidate if a specialization of $P$ is needed). For an adaptive system, such a decision is only based on available evidence about the two statements, represented by their uncertainty.

This problem is similar to the decision-making problem studied by the Bayesian school, where the system simply takes the option that has a higher probability (when they have the same utility). What makes things complex in NARS is the fact that here the uncertainty of a statement is represented by a *pair* of real numbers, and both numbers influence the system's preferences, but in different ways. When the competing statements have the same confidence, the system takes the one with a higher frequency as more likely to be confirmed. When the competing statements have the same frequency, the statement with a higher confidence is "stronger." For example, if $f_1 = f_2 = 1$, $c_1 = 1/2$, and $c_2 = 10/11$, the second statement is stronger, because it is supported by more evidence. In this case, the one with a higher confidence is more likely to be confirmed in the future. On the contrary, if $f_1 = f_2 = 0$, the one with a higher confidence is still "stronger", but less likely to be confirmed, because it has more negative evidence.

In general, we need to combine $f$ and $c$ into a single measurement $e$, indicating the system's *expectation* on how likely the statement will be confirmed again in the future. Intuitively, this measurement is similar to "probability" under subjective interpretation, which is derived from preference in decision making [1]. In other words, $e$ indicates the system's betting quotient on the statement, when the only alternative is the negation of the statement (that is, abstention is not allowed). To avoid a sure lose ("Dutch book"), the $e$ value of a statement and the $e$ value of the negation of the statement should sum to 1.

As defined previously, $c = 0$ means no evidence, therefore $e$ is $1/2$, since the system is indifferent to the statement and its negation. When $c = 1$, the system has known the limit of frequency, which is used as $e$. In other cases, $f$ is "squashed" by $c$ to the "indifference" point $1/2$ to become $e$, showing a "conservative" tendency by taking the possible variations of $f$ into consideration. Consequently, we obtain

$$e = c\,(f - 1/2) + 1/2.$$

When representing $e$ directly as a function of the weight of evidence, we get

$$e = \frac{w^+ + k/2}{w + k}$$

where $k$ is the constant defined previously. With the same evidence, a system with a larger $k$ has an $e$ closer to the indifference point, that is, it accepts a smaller betting quotient — the system is more prudent than a system with a smaller $k$. We call the $k$ a "personality parameter," because it shows a systematic bias in the system's preferences. Everyone prefers an option that has both a high frequency and a high confidence. However, when the two qualities cannot be achieved at the same time (i.e., one option has a higher frequency, but the other one has a higher confidence), different people balance the two differently. There is no "optimum value" for this parameter, as far as our current discussion concerns.

The expectation value also happen to be the middle point of the frequency interval

$$[\frac{w^+}{w + k}, \frac{w^+ + k}{w + k}]$$

### 3.3 Inferences

The major inference rules in NARS are the syllogistic rules for deduction, abduction, and induction, listed in the following. Each rule takes a pair of premises that share a common term, and derives a conclusion between the other two terms. Each rule includes a truth value function calculating the uncertainty of the conclusion from those of the premises.

**Deduction**

$$\begin{array}{l} M \subset P \; <f_1,\, c_1> \\ S \subset M \; <f_2,\, c_2> \\ \hline S \subset P \; <f,\, c> \end{array}$$

**Abduction**

$$\begin{array}{l} P \subset M \; <f_1,\, c_1> \\ S \subset M \; <f_2,\, c_2> \\ \hline S \subset P \; <f,\, c> \end{array}$$

**Induction**

$$\begin{array}{l} M \subset P \; <f_1,\, c_1> \\ M \subset S \; <f_2,\, c_2> \\ \hline S \subset P \; <f,\, c> \end{array}$$

A detailed discussion of the rules is beyond the scope of this paper, and such discussions can be found in [25, 26]. In the following, we only summarize the procedure by which the above functions is determined.

By definition, frequency $f$ and confidence $c$ take their values in the interval $[0, 1]$, and so does the amount of evidence $w$ when the evidence under consideration is at most of unit amount. Under this condition, we can carry out the task in the following steps:

1. Treat all the involved variables as Boolean, that is, have values in $\{0, 1\}$ (i.e., either 0 or 1). Consequently, each premise is fully positive ($f = 1,\ c = 1$), fully negative ($f = 0,\ c = 1$), or fully unknown ($c = 0$).

2. Study each value combination of the premises, and decide the corresponding values for the conclusion according to the semantics of the language and the definition of the uncertainty measurements.

For deduction, the Boolean truth value function is given by the transitivity of ideal Inheritance and the principle that from two pure negative Inheritance relations, no conclusion can be derived. There are the following situations:

- When $<f_1, c_1>$ and $<f_2, c_2>$ are both $<1, 1>$, so is $<f, c>$.
- When $<f_1, c_1>$ and $<f_2, c_2>$ are $<1, 1>$ and $<0, 1>$ (no matter which is which), $<f, c>$ is $<0, 1>$.
- When $<f_1, c_1>$ and $<f_2, c_2>$ are both $<0, 1>$, $c$ is 0.
- When $c_1$ or $c_2$ is 0, $c$ is 0.

For abduction and induction, the confidence of the conclusion cannot be 1, therefore it is not fruitful to directly represent $<f, c>$ as Boolean function of the truth values of the premises. Instead, the previous definition of evidence is used, so that the amount of evidence of the conclusion is represented as Boolean function:

- When $<f_1, c_1>$ and $<f_2, c_2>$ are both $<1, 1>$, $M$ is positive evidence, i.e., $w = w^+ = 1$.
- When in abduction $<f_1, c_1>$ is $<1, 1>$ and $<f_2, c_2>$ is $<0, 1>$, or in induction $<f_1, c_1>$ is $<0, 1>$ and $<f_2, c_2>$ is $<1, 1>$, $M$ is negative evidence, i.e., $w = 1$ and $w^+ = 0$.
- When in abduction $<f_1, c_1>$ is $<0, 1>$, or in induction $<f_2, c_2>$ is $<0, 1>$, $M$ is no evidence, i.e.,$w = 0$.
- When $c_1$ or $c_2$ is 0, $M$ is no evidence, i.e.,$w = 0$.

3. Represent the uncertainty of the conclusion as Boolean functions of those of the premises, under the constraint provided by the previous step. Usually there is more than one function satisfying the requirement, and we use the one that is simple and has a natural interpretation. What we get are:

**Deduction**
$AND(f, c) = AND(f_1, c_1, f_2, c_2)$
$c = AND(c_1, c_2, OR(f_1, f_2))$

**Abduction**
$w^+ = AND(f_1, c_1, f_2, c_2)$
$w = AND(f_1, c_1, c_2)$

**Induction**
$w^+ = AND(f_1, c_1, f_2, c_2)$
$w = AND(c_1, f_2, c_2)$

4. Extend the Boolean operators $AND$, $OR$, and $NOT$ from {0, 1} to [0, 1], according to the study of *T-norm* and *T-conorm* [2, 6, 17]. The extended Boolean operators in NARS are:

$$\begin{array}{lcl} AND(x, y) & = & x * y \\ OR(x, y) & = & x + y - x * y \\ NOT(x) & = & 1 - x \end{array}$$

where the first two are applied only when $x$ and $y$ are independent to each other, meaning that the value of one provides no information on the value of the other. When these operators are applied to the truth value functions obtained previously, we get the truth value function of NARS:

**Deduction**
$f = f_1 f_2/(f_1 + f_2 - f_1 f_2)$
$c = c_1 c_2(f_1 + f_2 - f_1 f_2)$

**Abduction**
$f = f_2$
$c = f_1 c_1 c_2/(f_1 c_1 c_2 + 1)$

**Induction**
$f = f_1$
$c = f_2 c_1 c_2/(f_2 c_1 c_2 + 1)$

The above inference rules, plus the revision rule and the choice rule introduced previously, are the major operations on confidence in NARS. By comparing them, we can see the following:

- Both frequency and confidence contribute to inference and decision making, but in different ways.
- Revision is the only rule where the confidence of the conclusion may be higher than those of the premises.
- The confidence of a syllogistic conclusion is never higher than the confidence of either premise, that is, confidence "declines" in syllogistic inference.
- Confidence declines much slower in deduction than in induction and abduction. In deduction, if both premises have a confidence value of 1, the conclusion may also have a confidence value of 1 (so it is a derived definition or convention). In induction and abduction, on the contrary, the confidence of the conclusion has an upper bound which is far less than 1. So, by saying that "Induction and abduction are more uncertain when compared with deduction", what is referred to is not the "first-order uncertainty", $f$ (inductive and abductive conclusions can have a frequency of 1 when all available evidence is positive), but the "higher-order uncertainty", $c$.

## 4 Compared with Other Approaches

### 4.1 Bayesian approach

For reasoning under uncertainty, the most popular research paradigm is the Bayesian approach, which has the following major features: [1, 4, 14, 20]

1. The *probability* of a proposition is interpreted as the system's degree of belief on the proposition, according to available evidence.
2. The system's beliefs, or knowledge, are represented by a (consistent) probability distribution on a proposition space.
3. When the system needs make a choice among competing uncertain answers, it always prefer the one that has the highest probability (when utility is the same).
4. The inferences in the proposition space precisely follow probability theory.
5. When new evidence comes, the beliefs are revised according to Bayes' theorem,

According to this approach, when "probability" is identified with "degree of belief", which indicates a system's preference among possible choices, and probability theory is used as a normative theory for how the system should behave to maintain a consistent belief space, a probability distribution on a proposition space is capable of representing the uncertainty involved in the above operations, because no matter what is the *origin* of uncertainty, its *effects* eventually appears in the system's preference among possible options in making a choice.

This argument is valid if we only consider the choice and the inference operations defined above. However, if we carefully analyze the revision operation, a limit of Bayesian approach can be found. A detailed discussion of this issue is in [23], and here we only briefly summarize the argument.

In the Bayesian approach, learning of new evidence is mainly carried out by *conditionalization* according to Bayesian Theorem, that is, if new evidence is $E$, then the probability of an arbitrary statement $S$ is changed from time $T0$ to time $T1$ as $P_{T1}(S) = P_{T0}(S|E)$. The problem in this method is: the knowledge that can be put into the system *a priori* (in $P_{T0}$) cannot always be learned *a posteriori* as $E$. In general, to revise the background knowledge of a probability distribution, to know the distribution itself is not enough. The confidence measurement introduced in this paper can be seen as an attempt to measure the knowledge behind a probabilistic judgment.

To use two (or more) numbers to represent the uncertainty of a statement is not a new idea. The previous problem (or similar problems) is the origin of many alternative approaches which challenge the dominant position of the Bayesian approach in the field. The advocates of these new approaches claim that the Bayesian approach cannot properly represent and process this kind of uncertainty, and various new measurements have been proposed. Some of them are discussed in the following.

### 4.2 Higher-order probability

Several new measurements are proposed under the assumption that the first-order uncertainty measurement (call it "probability" or "degree of belief") is an *approximation* of a "real" or "objective" probability. Under the frequentist interpretation, the probability of a statement is the limit of frequency, therefore all estimations of it based on finite evidence are not accurate. Even if we take probability as degree of belief, it can still be argued that such a degree should converge to the objective probability if it exists.

If the first-order probability assignment is only an approximation of an unknown value, the need for a higher-order measurement follows naturally — we want to know how good the approximation is, in addition to the approximated value itself.

One natural idea is to apply probability theory once again, which leads to the concepts like "probability of probability," "second order probability," "higher order probability," and so on [7, 8, 13]. In this way, we can assign probability to a probability assignment, to represent how good an approximation it is to the real probability.

However, there are problems in how to interpret the second value, and whether it is really useful [12, 14]. For our current purpose, under the assumption of insufficient knowledge, it makes little sense to talk about the "probability" that "the frequency is an accurate estimate of an objective first-order probability". Since NARS is always open to new evidence, it is simply impossible to decide whether the frequency of a judgment will converge to a point in the infinite future, not to mention where the point will be. If we say that the second-order probability is an approximation itself, then a third-order probability follows for the same reason — we are facing an infinite regression [16].

Since second-order probability is not introduced as a function of available evidence alone, it does not represent *ignorance*. "$P(P(A) = p) = 0$" means

"$P(A) \neq p$", rather than "$P(A)$ is completely unknown" (as what confidence means in NARS). Consequently, such a measurement does not support the revision operation — we cannot combine a pair of conflicting judgments, given their first-order and second-order probabilities.

Though the confidence value defined in NARS is in [0, 1], can be considered to be a ratio, and is at a "higher level" than $f$ (which is closely related to probability), in the sense that it indicates the stability of $f$, it cannot be interpreted as a second-order probability in the sense that it is the probability of the judgment "$f$ is the (real or objective) probability of the statement". The higher the confidence is, the harder it will be for the frequency to be changed by new evidence, but this does not mean that the judgment is "truer", or "more accurate", because in an open system like NARS, the concept of a real or objective probability does not exist.

Defined in this way, there is no "third-order uncertainty" to worry about. The stability of a confidence value can be derived from the confidence value itself. Because the current confidence is $c = w/(w + k)$, in the near future, with the coming of new evidence whose amount is $k$, the new confidence will be $(w + k)/(w + 2k)$, that is, $1/(2 - c)$. Therefore we do not need another measurement, and there is no infinite regression.

### 4.3 Probability Interval

Another intuitively appealing approach is to use an *interval*, rather than a *point*, to represent uncertainty, and to interpret the interval as the lower bound and upper bound of the real probability [2, 9, 11]. In this way, the higher-order uncertainty can be represented by the width of the interval. When the system knows nothing about a statement, the interval is [0, 1], so the real probability can be anywhere; when the real probability is known, the interval degenerates into a point.

A well-known method in statistics is to calculate the *confidence interval* [1], which has a high probability (such as 95%) of including within the real probability. Here, the width of the interval also provides information about the accuracy of the current estimation.

Although the above methods are directly based on probability theory (and thus have a sound foundation), and they are useful for various purposes, they cannot be applied in a system like NARS.

When the probability interval is interpreted as the interval containing the true probability value, the situation is similar to the case of higher-order probability. For an open system with insufficient knowledge, it cannot be assumed that a frequency always has a limit. Even when such a limit really exists, it is impossible for the system to know how close the current frequency is to it without making assumptions about the distribution of the limit. If the probability interval is interpreted as an estimation itself, an "interval of the bounds" will follow, so the infinite regression appear again. For the same reason, the "confidence" defined in NARS has different meaning from the "confidence" as in "confidence interval", used in probability theory and statistics, though they do correspond to the same intuition, that is, some frequency estimations are more reliable than the others, and their difference can be measured.

The closest probability-based approach to NARS is the "imprecise probability" (IP) theory proposed by Peter Walley [21, 22]. Walley defines lower and upper probabilities of an event as the minimum and maximum betting rate, respectively, that a rational person is willing to pay for the gamble on the event. An interesting result relates this theory to the approach proposed in this paper. Suppose that an event has a constant (unknown) chance to happen, that the observations of the event are independent to one another, and that the chance has a *near-ignorance* beta distribution as its prior. If the observed relative frequency of the event is $m/n$, then, according to Walley's theory, the lower and upper probabilities of the event are $m/(n + s_0)$ and $(m + s_0)/(n + s_0)$, respectively. Here $s_0$ is a parameter of the beta distribution, and it indicates the convergence speed of the lower and upper probabilities. This is exactly the result we get for the lower and upper frequencies previously. The width of the interval can be seen as a measurement of "ignorance", which is the opposite of "confidence" defined above (the sum of the two is 1).

Though these two approaches (NARS and IP) define uncertainty measurements differently, they are consistent in the sense that they make the same decisions in situations where both theories are applicable. What makes them different from the other "probability interval" approaches mentioned earlier is: in both NARS and IP, the interval does not bound the limit of the frequency (if such a limit exists). The interval $[m/(n + s_0),\ (m + s_0)/(n + s_0)]$ is just where the frequency will be in a *constant near future*, and after that it can be anywhere in [0, 1].

The major difference between these two approaches comes from the fact that IP is proposed as an extension of probability theory, and therefore the inference is mainly within the same probability distribution. On the other hand, NARS is designed to be a *logic*. As described previously, in NARS each piece of knowledge is based on a separate body of evidence, so that the rules introduced previously correspond to inference across different probability distributions. The detailed relationship between these two approaches is an interesting issue for future research.

### 4.4 Other related works

The Dempster-Shafer theory [5, 18] is also motivated by the observation that probability theory cannot represent and process ignorance properly. However, the use of Dempster's rule causes a conflict between the two basic goals of the theory — that is, generalizing probability theory and supporting evidence combination. This problem is discussed in [24], and thus is not repeated in this paper.

The expectation continuum in NARS

$$e = \frac{w^+ + k/2}{w + k}$$

is also very similar to Carnap's $\lambda$-continuum [3] and the result derived in probability theory for a *beta distribution* [1]. Again, what make our formula different from them is that it is completely based on available evidence without any reference to a "real probability". As we have shown, such a definition is necessary for the purpose of revision.

Intuitively, confidence indicates how much the system knows about a statement, and thus is similar to Shafer's "reliability" [19] or Yager's "credibility" [28]. Both of the approaches evaluate the uncertainty of a probability assignment, where 0 is interpreted as "unknown", rather than "impossible". These approaches relate the higher-order measurement to the reliability of its information source or to the compatibility of a judgment with higher priority evidence. These kinds of information, though available in some other situations, is not available in NARS, where the system's confidence about a statement depends on the amount of available evidence.

## 5 Summary

With conflicting evidence, the conclusions made by a reasoning system are usually uncertain. If the system is also open to new evidence, there is a higher order uncertainty indicating the stability of first order uncertainty evaluations.

Several approaches have been proposed for handling higher order uncertainty. Although each approach has a suitable application domain, none is appropriate for the situation discussed in this paper.

When positive and negative evidence can be (in principle) distinguished and measured, *confidence* can be defined as a function of the amount of available evidence. Such a definition is simple, natural, and closely related to concepts like ignorance, credibility, reliability, stability, sensitivity, susceptibility, and so on. This measurement also provides support to the design of various reasoning rules.

Though the uncertainty calculus used in NARS shares certain intuition and even concrete formula with probability-based approaches, it is not an application or extension of probability theory, for several reasons.

The frequency and confidence measurement used in NARS is completely defined on available evidence, without any assumption about the distribution, or even the existence, of the limit of the frequency.

When measuring frequency and confidence in NARS, both extensional and intentional factors are included (as defined previously). Therefore, the frequency of "$A \subset B$" is not the conditional probability $P(x \in E_B \mid x \in E_A) = |E_B \cap E_A| \;/\; |E_A|$ ("The probability for $A$'s instance to be included in $B$"), which is the usual (pure extensional) interpretation of probability.

In NARS, the uncertainty of each statement is evaluated separately, based on local evidence, so the system's beliefs do not correspond to a consistent probability distribution over the space of all statements. Instead, each statement is like a probability distribution of its own, with its statement space and background knowledge.

Consequently, the uncertainty calculations carried out by the inference rules do not correspond to probability calculations within the same probability distribution. They are more similar to calculations across multiple probability distributions.

The NARS approach is not proposed as heuristics or an ad hoc method for special purposes. Though still incomplete, this approach is designed to be a normative theory of inference. Its differences with probability theory mainly come from its assumption of insufficient knowledge and resources [27].

This approach is not necessarily better than the competing approaches in all environments, but it is better in the environment described at the beginning of the paper, which has special theoretical and practical interests from the view point of artificial intelligence and cognitive science [26].

## References


[1] J. Bernardo and A. Smith. *Bayesian Theory.* John Wiley & Sons, Chichester, England, 1994.

[2] P. Bonissone. Summarizing and propagating un-

certain information with triangular norms. *International Journal of Approximate Reasoning*, 1:71–101, 1987.

[3] R. Carnap. *The Continuum of Inductive Methods.* The University of Chicago Press, Chicago, 1952.

[4] P. Cheeseman. In defense of probability. In *Proceedings of the Eighth International Joint Conference on Artificial Intelligence*, pages 1002–1009, 1985.

[5] A. Dempster. Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, 38:325–339, 1967.

[6] D. Dubois and H. Prade. A class of fuzzy measures based on triangular norms. *International Journal of General Systems*, 8:43–61, 1982.

[7] R. Fung and C. Chong. Metaprobability and Dempster-Shafer in evidential reasoning. In L. Kanal and J. Lemmer, editors, *Uncertainty in Artificial Intelligence*, pages 295–302. North-Holland, Amsterdam, 1986.

[8] H. Gaifman. A theory of higher order probabilities. In J. Halpern, editor, *Theoretical Aspects of Reasoning about Knowledge*, pages 275–292. Morgan Kaufmann, Los Altos, California, 1986.

[9] B. Grosof. An inequality paradigm for probabilistic knowledge. In L. Kanal and J. Lemmer, editors, *Uncertainty in Artificial Intelligence*, pages 259–275. North-Holland, Amsterdam, 1986.

[10] J. Keynes. *A Treatise on Probability.* Macmillan, London, 1921.

[11] H. Kyburg. Bayesian and non-Bayesian evidential updating. *Artificial Intelligence*, 31:271–293, 1987.

[12] H. Kyburg. Higher order probabilities. In L. Kanal, T. Levitt, and J. Lemmer, editors, *Uncertainty in Artificial Intelligence 3*, pages 15–22. North-Holland, Amsterdam, 1989.

[13] G. Paaß. Second order probabilities for uncertain and conflicting evidence. In P. Bonissone, M. Henrion, L. Kanal, and J. Lemmer, editors, *Uncertainty in Artificial Intelligence 6*, pages 447–456. North-Holland, Amsterdam, 1991.

[14] J. Pearl. *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann Publishers, San Mateo, California, 1988.

[15] C. Peirce. The probability of induction. *Popular Science Monthly*, 1878. Reprinted in *The World of Mathematics 2*, J. Newman, ed., New York: Simon and Schuster (1956), 1341–1354.

[16] L. Savage. *The Foundations of Statistics.* Wiley, New York, 1954.

[17] B. Schweizer and A. Sklar. *Probabilistic Metric Spaces.* North-Holland, Amsterdam, 1983.

[18] G. Shafer. *A Mathematical Theory of Evidence.* Princeton University Press, Princeton, New Jersey, 1976.

[19] G. Shafer. Perspectives on the theory and practice of belief functions. *International Journal of Approximate Reasoning*, 4:323–362, 1990.

[20] D. Spiegelhalter. A statistical view of uncertainty in expert systems. In W. Gale, editor, *Artificial Intelligence and Statistics*, pages 17–56. Addison Wesley, Reading, 1986.

[21] P. Walley. *Statistical Reasoning with Imprecise Probabilities.* Chapman and Hall, London, 1991.

[22] P. Walley. Inferences from multinomial data: learning about a bag of marbles. *Journal of the Royal Statistical Society, Series B*, 58:3–57, 1996.

[23] P. Wang. Belief revision in probability theory. In *Proceedings of the Ninth Conference on Uncertainty in Artificial Intelligence*, pages 519–526. Morgan Kaufmann Publishers, San Mateo, California, 1993.

[24] P. Wang. A defect in Dempster-Shafer theory. In *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*, pages 560–566. Morgan Kaufmann Publishers, San Mateo, California, 1994.

[25] P. Wang. From inheritance relation to nonaxiomatic logic. *International Journal of Approximate Reasoning*, 11(4):281–319, November 1994.

[26] P. Wang. *Non-Axiomatic Reasoning System: Exploring the Essence of Intelligence.* PhD thesis, Indiana University, 1995.

[27] P. Wang. Heuristics and Normative Models of Judgment Under Uncertainty. *International Journal of Approximate Reasoning*, 14(4):221–235, May 1996.

[28] R. Yager. Credibility discounting in the theory of approximate reasoning. In P. Bonissone, M. Henrion, L. Kanal, and J. Lemmer, editors, *Uncertainty in Artificial Intelligence 6*, pages 299–310. North-Holland, Amsterdam, 1991.