Probability Theory Ensues from Assumptions of Approximate Consistency: A Simple Derivation and its Implications for AGI

Ben Goertzel

Novamente LLC

Abstract. By slightly tweaking some recent mathematics by Dupré and Tipler, it is shown that, if an intelligence even *approximately* obeys certain simple consistency conditions in its reasoning about uncertainty, then its uncertainty judgments must *approximately* obey the rules of probability theory. It is argued that, while real-world cognitive systems will rarely be fully consistent, they will often possess approximate consistency of the sort required by this mathematics. The conclusion is that, much of the time, cognitive systems' uncertainty judgments will roughly resemble the conclusions of probability theory – with more deviation from probability theory when fewer resources are devoted to the maintenance of consistency. Implications for AGI architecture are briefly discussed, including the viability of hybrid systems that use (explicit or implicit) probabilistic inference when resources are available for adequate consistency maintenance, and other heuristic methods elsewise.

1 Introduction

Is probability theory *the* correct way to reason about uncertainty? Or is it just one tool among many, to be considered alongside options like possibility theory [**?**], NARS [**?**], Dempster-Shafer theory [**?**] and so forth?

The philosophical foundations of probability theory are surprisingly complex [**?**], and simplistic "frequentist" interpretations of probability don't hold up well. However, axiomatic approaches are widely seen as more compelling. In our view, the strongest argument that probability theory is the uniquely right approach to uncertain reasoning rests on various derivations of the rules of probability from sets of axioms that appear intuitively evident.

Cox's classic work [**?**] from the middle of the last century set the tone for this line of research, deriving the rules of probability from mathematical versions of three simple principles:

- plausibilities are to be represented as real numbers (lifting this principle but keeping the others leads to Youssef's theory of exotic probabilities [**?**], which may take complex, quaternionic or octonionic values)
- plausibilities should be commonsensical, including agreement with standard Boolean logic in the case of statements that are completely plausible or completely implausible

– plausibilities are to be consistent, in the sense that if there are multiple sensible ways to calculate a certain plausibility, they should all yield the same answer

Based on these conceptual principles, Cox proposed the axioms

1. Where f is the function mapping a proposition's plausibility into the plausibility of its negation, $f(f(x)) = x$
2. The plausibility of the conjunction $A\&B$ depends only on the plausibility of B and on the plausibility of A given B is true.
3. Suppose $A\&B$ is equivalent to $C\&D$. Then the following two scenarios lead to the same result: 1) acquire new information A, then acquire more information B and update the probabilities; 2) acquire new information C, then acquire more information D and update the probabilities.

From these axioms, Cox derived the conclusion that the plausibility must be a monotonically scaled version of probability, obeying the usual probabilistic rules on finite sets. The third axiom encapsulates the conceptual requirement of "consistency."

Cox's work was classic, yet had its shortcomings. Cox's derivation was not entirely rigorous, and it relied on additional assumptions regarding the smoothness of uncertainty measures. Halpern reported technical counterexamples [**?**] [**?**]. Several recent papers have presented novel arguments in the spirit of Cox's, with weaker assumptions and more rigorous proofs [**?**] [**?**] [**?**]. Dupré and Tipler [**?**] have presented a particularly simple set of Cox-like axioms and shown that the rules of probability follow from these, though it's arguable that the assumptions used by Knuth and Skilling [**?**] are both weaker and more elegant.

A troublesome aspect of the original Cox approach, from an AGI and cognitive science perspective, is that Cox's third axiom intuitively appears too strong to be literally applied to real-world complex cognitive systems. Human beings do not display perfect Cox-type consistency, and nor would any complex real-world AGI system confronted with a large volume of data to reason about. The same core conceptual problem is faced by the more sophisticated descendants of Cox's approach, referenced above. This raises the question whether the correctness of probability theory is restricted to unrealistic "toy" or formalized situations where reasoning systems can afford to operate with perfect consistency.

One way to resolve this question would be to show that, if Cox-type consistency holds *approximately*, then uncertain inference must obey the rules of probability with an appropriate degree of approximation. This would connect the axiomatic foundation of probability theory more closely with the operation of real-world cognitive systems. Real human minds and AGI systems can never be perfectly consistent, but in many situations they can approximate consistency.

In this paper we exploit the extreme simplicity of Dupré and Tipler's axiomatic foundation for probability theory, to give a very simple demonstration that approximate Cox-type consistency implies approximate adherence to the

rules of probability. Or in a phrase that was almost used to title this paper: *Probably Approximately Consistent Plausibilities are Probably Approximately Probabilities.*

We then explore the conceptual implications of this result for AGI and cognitive science. We suggest that, in circumstances where a cognitive system can afford to devote the resources to achieve a reasonable degree of Cox-type consistency, it will then reason about uncertainty using probability theory. On the other hand, when resources for inference about a given topic are too scant to ensure reasonable consistency, other heuristics may sensibly be used.

The mathematical portion of this paper relies on Dupré and Tipler's formalization, merely because it is the simplest of the many "Cox-like" axiomatic derivations of probability theory. It seems likely that similar derivations can be performed for other Cox-like derivations as well, e.g. the approach in [**?**] which seems particularly powerful.

2 Dupré and Tipler's Axiomatic Derivation of Probability Theory

Dupré and Tipler [**?**] derive the rules of probability theory from a few simple axioms. Here we merely state the axioms without explanation. However, the reader is directed to Dupré and Tipler's longer, earlier paper [**?**] for an intuitive explanation of the meaning of the axioms and why they are cognitively natural. Note that the "unknown numbers" used in their formalism are essentially analogous to random variables in the conventional Kolmogorov formulation of probability theory. The mathematical essence of their derivation of probability theory is based on the properties of retraction mappings from unknown numbers into known ones.

- **AXIOM 1. STRUCTURE OF UNKNOWN REAL NUMBERS AND PLAUSIBLE VALUE.** [1] We assume a set T of unknown numbers is a partially ordered commutative algebra over the real numbers R with identity, 1.
  - In addition, we assume a given sub-Boolean algebra E of $E(T)$ with $0, 1 \in E$ and denote by E_0 the set of non-zero members of E. We assume that the partial ordering in $E(T)$ as a Boolean algebra coincides with the ordering that $E(T)$ inherits from the algebra T. Finally, we assume a function $PV : T \times E_0 \to R$, called PLAUSIBLE VALUE, whose value on the pair (x, e) is denoted $PV(x|e)$.
- **AXIOM 2. STRONG RESCALING FOR PLAUSIBLE VALUE.** If a, b belong to R, if x belongs to T, and if e belongs to E_0, then $PV(ax+b|e) = aPV(x|e) + b$.
- **AXIOM 3. ORDER CONSISTENCY FOR PLAUSIBLE VALUE.** If $x, y \in T$ and if $e \in E_0$, implies that $x \leq y$, then $PV(x|e) \leq PV(y|e)$.

[1] These axioms are quoted, with slight textual modifications, from [**?**]

– **AXIOM 4. THE COX AXIOM FOR PLAUSIBLE VALUE** : If e, c are fixed in E, with $ec \in E_0$, if x_1, x_2 are in T, if $PV(x_1|ec) = PV(x_2|ec)$, then $PV(x_1e|c) = PV(x_2e|c)$. That is, we assume that as a function of x, the plausible value $PV(xe|c)$ depends only on $PV(x|ec)$.
– **AXIOM 5. RESTRICTED ADDITIVITY OF PLAUSIBLE VALUE** . For each fixed $y \in T$ and $e \in E_0$,the plausible value $PV(x+y|e)$ as a function of $x \in T$ depends only on $PV(x|e)$, which is to say that if $x_1, x_2 \in T$ and $PV(x_1|e) = PV(x_2|e)$, then $PV(x_1 + y|e) = PV(x_2 + y|e)$.

Using these axioms, and writing $PL(A, B) = PV(A, B)$ for $A \in E$, $B \in E_0$ by regarding $E \subset E(T)$, they arrive at the conclusion

Theorem 1. *(Dupré and Tipler). If e, c, g belong to E, with $ec \in E_0$, then*

$$PL(gc|e) = PL(g|ce)PL(c|e)$$

which is the product rule that is the outcome of Cox's original argument. The rest of Cox's argument for probability rules follows in a familiar way. For instance, Theorem 1 together with AXIOMS 1 and 3 yield the disjunction rule

$$PL(AorB|C) = PL(A|C) + PL(B|C) - PL(A\&B|C)$$

And AXIOM 5 can be used to prove finite additivity of plausibility.

3 Approximate Consistency Implies Approximate Adherence to Probability Rules

Due to the impressive simplicity of Dupré and Tipler's argument, it is almost trivial to modify it to yield a proof that approximate Cox-type consistency implies approximate adherence to the probability rules. More precisely: Suppose we assume Axioms 1-3 and 5 from the Dupré and Tipler argument, as summarized above, but weaken the consistency axiom to:

– **AXIOM 4′ . APPROXIMATE COX AXIOM FOR PLAUSIBLE VALUE** : Assume a given probability distribution μ on E, and another distribution ν on T. Assume $e, c \in E$ are chosen from μ, subject to the constraint $ec \in E_0$; and assume $x_1, x_2 \in T$ are drawn from ν. Then; if $PV(x_1|ec) = PV(x_2|ec)$,

$$P(|PV(x_1e|c) - PV(x_2e|c)| < \epsilon) > 1 - \delta$$

That is, we assume that as a function of x, it is "probably approximately correct" that the plausible value $PV(xe|c)$ depends entirely on $PV(x|ec)$. Here ϵ and δ are fixed numbers (and of course the axiom is more interesting if they are reasonably small.)

Then the conclusion we find is

Corollary 1. *If e, c, g are drawn from μ on E, subject to the constraint $ec \in E_0$, then*

$$P(|PL(gc|e) - PL(g|ce)PL(c|e)| < 2\epsilon) > 1 - 2\delta$$

Mirroring Dupré and Tipler's proof of Theorem 1, we obtain this as an immediate corollary of

Theorem 2. *Assume axioms 1-3 and 4′ If $x \in T$, is drawn from ν, and if $e, c \in E$ are drawn from μ subject to $ec \in E_0$, then $|PV(xc|e) - PV(x|ce)PV(c|e)| = |PV(xc|e) - PV(x|ce)PL(c|e)|$, and*

$$P(|PV(xc|e) - PV(x|ce)PL(c|e)| < 2\epsilon) > 1 - 2\delta$$

We prove this, following Dupré and Tipler's proof of their corresponding lemma, as follows:

Proof. We consider a real-valued function

$$F = F(c, e) : T \to RF(x) = PV(xc|e).$$

defined by

$$F(x) = PV(xc|e)$$

AXIOM 4' says that – within an error of ϵ – $F(x)$ depends only on the numerical value $PV(x|ce)$ and not on the particular $x \in T$. Put differently, it implies that there is a real-valued function $f = f(c, e) : R \to R$ so that

$$P(|F(x) - f(PV(x|ce))| < \epsilon) > 1 - \delta$$

If we take the case of $x = r \in R \subset T$, then $F(x) = F(r) = PV(rc|e) = rPV(c|e)$, by AXIOM 5 above. But as $r = PV(r|ce)$, this means,

$$\begin{aligned} |f(r) - rPV(c|e)| = \\ |f(r) - F(r) + F(r) - rPV(c|e)| \leq \\ |f(r) - F(r)| + |F(r) - rPV(c|e)| \\ < \epsilon + 0 = \epsilon \end{aligned}$$

for all $r \in R$, which in turn means f is – in a "probably approximately correct" sense – simply multiplication by $PV(c|e) = PL(c|e)$. But now, if $x \in T$ is arbitrary, we conclude that, with probability $> 1 - 2\delta$,

$$\begin{aligned} |PV(xc|e) - PV(x|ce)PL(c|e)| = \\ |PV(xc|e) - f(PV(x|ce)) - f(PV(x|ce)) - PV(x|ce)PL(c|e)| \leq \\ |PV(xc|e) - f(PV(x|ce)| + |f(PV(x|ce)) - PV(x|ce)PL(c|e)| \leq 2\epsilon \end{aligned}$$

QED

The same method may be used to prove probably approximately correct additivity of the plausibility, based on a modified AXIOM 5 such as

– **AXIOM 5′. APPROXIMATE RESTRICTED ADDITIVITY OF PLAUSIBLE VALUE** . Suppose $y \in T$ is drawn from probability distribution ν, and $e \in E_0$ is drawn from probability distribution μ. Then, the statement that the plausible value $PV(x + y|e)$ as a function of $x \in T$ depends only on $PV(x|e)$ is probably approximately correct, which is to say that if $PV(x_1|e) = PV(x_2|e)$, then

$$P(|PV(x_1 + y|e) - PV(x_2 + y|e)| < \epsilon) > 1 - \delta$$

for suitable, fixed ϵ and δ.

Qualitatively, the message is that: If Cox-type consistency approximately holds, then plausibilities behave approximately like probabilities. This is intuitively what one would expect, and it falls right out of Dupré and Tipler's formalism.

My suspicion is that a similar conclusion could be obtained regarding other axiomatic derivations of probability theory as well, but the mathematics would be considerably more difficult. The other derivations, while elegant, involve solutions to functional equations, which would then have to be re-calculated using interval arithmetic or some other method of dealing with an approximative consistency assumption. This is an interesting area for future research.

4 Implications for Cognitive Systems

Do these considerations finally resolve the matter once and for all, and assert probability as *the* correct way for cognitive systems to measure uncertainty?

Not quite. What they do is make clear that: When a cognitive system has sufficient resources to be reasonably (not necessarily perfectly) consistent in thinking about a certain topic, probability theory is the correct way to manage uncertainty regarding that topic.

However, the best option for a cognitive system when it can't devote enough resources to avoid massive inconsistency regarding some topic, remains unclear. Should the system try to emulate probability theory as best it can? Or is there some other heuristic that works better in this kind of situation?

The philosophy underlying Pei Wang's NARS system [**?**], insofar as I understand it, appears to be that the maintenance of reasonable consistency is simply not feasible for real-world general intelligences, rendering probability theory irrelevant except in special cases. The NARS logic is then rated as applicable based on its adherence to different axioms, not including the sort of consistency required by the Cox axioms or similar approaches.

The Probabilistic Logic Networks [**?**] system that I have co-created and am currently working with, allows global inconsistency in the knowledge base it works with, but seeks to build local "islands of consistency" associated with

its current focus of reasoning. When it cannot afford to spend a lot of effort establishing consistency, it still applies probabilistic rules, doing the best it can to stay close to probability theory.

On the other hand, the overall OpenCog system in which PLN is embedded [**?**] contains other cognitive algorithms besides PLN, including a system called ECAN [**?**] that resembles attractor neural networks with Hebbian learning. ECAN and various other OpenCog cognitive dynamics don't care about inconsistency at all. It is possible that Hebbian learning type methods such as ECAN possess some sort of optimality for the low-consistency situation. So it is possible that the best solution for a real-world cognitive system is to use probabilistic inference for the focus of its attention, where it can maintain reasonable consistency, and use Hebbian-type methods for broader background inference where it cannot afford to maintain reasonable consistency. But this remains speculation, to be resolved via future work.

It seems that it might be possible to prove a theorem of the following vague form: If a cognitive system, given a certain amount of resources, wants to maximize the expected degree to which it can fulfill some goal over its future, then it will do best to allocate a significant percentage of its resources to maintaining Cox-type consistency regarding its knowledge pertaining to the goal and related actions and perceptions. Because if it does this, then its reasoning about goal-achievement will adhere approximately to probability theory, and its strategy for goal-achievement will be more likely to maximize the expected degree of goal-achievement. Of course, if a theorem like this works, the key will be that the ultimate objective of the cognitive system is being posed using probability theory (in terms of the expected degree of goal achievement). But at the present time, the exact form such a theorem might take remains unclear, so this is also speculation to be potentially resolved via future research.