Probability on Sentences in an Expressive Logic

Marcus Hutter **John Lloyd**
Kee-Siong Ng **Will Uther**

Canberra, ACT, 0200, Australia
http://www.hutter1.net/

THE AUSTRALIAN NATIONAL UNIVERSITY

2012

Motivation

- Automated reasoning about uncertain knowledge has many applications.
- One difficulty when developing such systems is the lack of a completely satisfactory integration of logic and probability.
- We address this problem head on.

Induction Example: Black Ravens

- Consider a sequence of ravens identified by positive integers.
- Let $B(i)$ denote the fact that raven i is black. $i = 1, 2, 3, ...$
- We see a lengthening sequence of black ravens.
- Consider the hypothesis "all ravens are black" $\widehat{=} \forall i.B(i)$:
- Intuition: Observation of black ravens with no counter-examples increases confidence in hypothesis.
- Plausible requirement on any inductive reasoning system: Probability$(\forall i.B(i) \,|\, B(1) \wedge ... \wedge B(n))$ tends to 1 for $n \to \infty$.
- Real-world problems are much more complex, but most systems fail already on this apparently simple example.
- E.g. Bayes/Laplace rule and Carnap's confirmation theory fail [RH11],
- but Solomonoff induction works [RH11].

Logic & Probability

- Logic&Structure: Expressive languages like higher-order logic are ideally suited for representing and reasoning about structured knowledge.
- Probability&Uncertainty: Uncertain knowledge can be modeled by assigning graded probabilities rather than binary truth-values to sentences.
- Combined: Probability over Sentences.

Main Aim (main technical problem considered)

Given a set of sentences, each having some probability of being true, what probability should be ascribed to other (query) sentences?

- Alternative (not considered): Probability inside Sentences. Treated previously by Lloyd&Ng&Uther (2008-2009).

Natural Wish List (among others)

The probability distribution should

(i) be consistent with the knowledge base,

(ii) allow for a consistent inference procedure and in particular

(iii) reduce to deductive logic in the limit of probabilities being 0 and 1,

(iv) allow (Bayesian) inductive reasoning and

(v) learning in the limit and in particular

(vi) allow to confirm universally quantified hypotheses=sentences.

Technical Requirements

This wish-list translates into the following technical requirements for a prior probability: It needs to be

(P) consistent with the standard axioms of Probability,

(CA) including Countable Additivity,

(C) non-dogmatic $\widehat{=}$ Cournot
$\widehat{=}$ zero probability means impossibility
$\widehat{=}$ whatever is not provably false is assigned probability larger than 0.

(G) separating $\widehat{=}$ Gaifman $\widehat{=}$ existence is always witnessed by terms
$\widehat{=}$ logical quantifiers over variables can be replaced by meta-logical quantification over terms.

Main Results

- Suitable formalization of all requirements.
- Proof that probabilities satisfying all our criteria exist.
- Explicit constructions of such probabilities.
- General characterizations of probabilities that satisfy some or all of the criteria.
- Various (counter) examples of (strong) (non)Cournot and/or Gaifman probabilities and (non)separating interpretations.

Achievement (unification of probability & logic & learning)

The results are a step towards a globally consistent and empirically satisfactory unification of probability and logic.

More: Partial Knowledge and Entropy

- We derive necessary and sufficient conditions for extending beliefs about finitely many sentences to suitable probabilities over all sentences.
- Seldom does knowledge induce a unique probability on all sentences.
- In this case it is natural to choose a probability that is least dogmatic or least biased.

We show that the probability of minimum entropy relative to some Gaifman and Cournot prior

(1) exists, and is
(2) consistent with our prior knowledge,
(3) minimally more informative,
(4) unique, and
(5) suitable for inductive inference.

Outlook: how to use and approximate the theory for autonomous reasoning agents.

On the Choice of Logic

- We use: simple type theory = higher-order logic,
  Henkin semantics, no description operator, countable alphabet.
- But: The major ideas work in many logics (e.g. first order).
- But: There are important and subtle pitfalls to be avoided.
- But: No time to dig deep enough in this talk for any of this to matter.
- Slides will abstract away from and gloss over the details of the used logic.
- Logical symbols & conventions: boolean operations $\top, \bot, \wedge, \vee, \rightarrow$,
  quantifiers $\forall x, \exists y$, abstraction λz, closed terms t, sentences φ, χ,
  formula $\psi(x)$ with a single free variable x,
  universal hypothesis/sentence $\forall x.\psi(x)$, ...

Probability on Sentences

Definition (probability on sentences)

A probability (on sentences) is a non-negative function $\mu : \mathcal{S} \to \mathbb{R}$ satisfying the following conditions:

- If φ is valid, then $\mu(\varphi) = 1$.
- If $\neg(\varphi \wedge \chi)$ is valid, then $\mu(\varphi \vee \chi) = \mu(\varphi) + \mu(\chi)$.
- Conditional probability: $\mu(\varphi|\chi) := \dfrac{\mu(\varphi \wedge \chi)}{\mu(\chi)}$.

- $\mu(\varphi)$ is the probability that φ is valid in the intended interpretation, or
- $\mu(\varphi)$ is the subjective probability held by an agent that sentence φ holds in the real world.
- No Countable Additivity (CA) for μ – all sentences are finite.

Probability on Interpretations

- $mod(\varphi)$:= Set of (Henkin) Interpretations in which φ is valid.
- $\mathcal{I} := mod(\top) =$ set of all (Henkin) interpretations.
- $\mathcal{B} := \sigma$-algebra generated by $\{mod(\varphi) : \varphi \in \mathcal{S}\}$

Definition (probability on interpretations)

A function $\mu^* : \mathcal{B} \to \mathbb{R}$ is a (CA) probability on σ-algebra $\mathcal{B}$ if $\mu^*(\emptyset) = 0$ and $\mu^*(\mathcal{I}) = 1$ and for all countable collections $\{A_i\}_{i \in I} \subset \mathcal{B}$ of pairwise disjoint sets with $\bigcup_{i \in I} A_i \in \mathcal{B}$ it holds that $\mu^*(\bigcup_{i \in I} A_i) = \sum_{i \in I} \mu^*(A_i)$.

Probability on: Sentences $\Leftrightarrow$ Interpretations

Probability on sentences μ $\iff$ a measure-theoretic probability distribution μ^* on sets of interpretations $\mathcal{I} \in \mathcal{B}$.

Proposition ($\mu \Rightarrow \mu^*$)

Let $\mu : \mathcal{S} \to \mathbb{R}$ be a probability on $\mathcal{S}$. Then there exists a unique probability $\mu^* : \mathcal{B} \to \mathbb{R}$ such that $\mu^*(mod(\varphi)) = \mu(\varphi)$, for each $\varphi \in \mathcal{S}$.

Proof uses compactness of class of Henkin interpretations $\mathcal{I}$ and Caratheodory's unique-extension theorem.

Proposition ($\mu^* \Rightarrow \mu$)

Let $\mu^* : \mathcal{B} \to [0,1]$ be a probability on $\mathcal{B}$. Define $\mu : \mathcal{S} \to \mathbb{R}$ by $\mu(\varphi) = \mu^*(mod(\varphi))$, for each $\varphi \in \mathcal{S}$. Then μ is a probability on $\mathcal{S}$.

Proof is trivial.

Separating Interpretations

- Black raven ctd: Intuition: $\{B(1), B(2), ...\}$ should imply $\forall x.B(x)$.
- Problem: This is not the case: There are non-standard models of the natural numbers in which $x = n$ is invalid for all $n = 1, 2, 3, ...$.
- Solution: Exclude such unwanted interpretations.
- Generalize $1, 2, 3, ...$ to "all terms t".

Definition (separating interpretation)

An interpretation I is *separating* iff for all formulas $\psi(x)$ the following holds:

If I is a model of $\exists x.\psi(x)$,

then there exists a closed term t such that I is a model of $\psi\{x/t\}$.

- Informally: existence is always witnessed by terms.
- $\widehat{mod}(\varphi)$:= Set of separating models of φ, and $\widehat{\mathcal{I}} = \widehat{mod}(\top)$.
- $\widehat{\mathcal{B}}$:= σ-algebra generated by $\{\widehat{mod}(\varphi) : \varphi \in \mathcal{S}\}$
- All $\widehat{mod}(\varphi)$ are $\mathcal{B}$-measurable.

Gaifman Condition

Effectively avoid non-separating interpretations
by requiring probability on them to be zero.

Definition (Gaifman condition)

$\mu(\forall x.\psi(x)) = \lim_{n\to\infty} \mu(\bigwedge_{i=1}^{n} \psi\{x/t_i\})$ for all ψ, where $t_1, t_2, ...$ is an enumeration of (representatives of) all closed terms (of same type as x).

Informally: logical quantifiers over variables can be replaced by meta-logical quantification over terms.

Theorem ($\mu^*(\mathcal{I} \setminus \widehat{\mathcal{I}}) = 0 \quad \Leftrightarrow \quad \mu$ is Gaifman)

The Gaifman condition (only) forces the measure of the set of non-separating interpretations to 0.

Induction Still does Not Work

- $\mu(\forall i.B(i) \,|\, B(1) \wedge ... \wedge B(n)) \equiv 0$ if $\mu(\forall i.B(i)) = 0$.
- This is the infamous Zero-Prior problem in philosophy of induction.
- Carnap's and most other confirmation theories fail,
  since they (implicitly & unintentionally) have $\mu(\forall i.B(i)) = 0$.
- Why is this problem hard?
  "Naturally" $\mu(\forall i.B(i)) \;\leq\; \mu(B(1) \wedge ... \wedge B(n)) \;\to\; 0$
  (Think of independent events with prob. $p < 1$, then $p \cdot p \cdot p \cdots \to 0$)
- But it's not hopeless:
  Just *demand* $\mu(\forall x.\psi(x)) > 0$ for all ψ for which this is possible.

Cournot Condition

Cournot's principle informally

- $\widehat{=}$ probability zero/one means impossibility/certainty
- $\widehat{=}$ whatever is not provably false is assigned probability larger than 0
- $\widehat{=}$ all (sensible) prior probabilities should be non-zero
- $\widehat{=}$ be as non-dogmatic as possible

Definition (Cournot probability)

A probability $\mu : \mathcal{S} \to \mathbb{R}$ is Cournot if, for each $\varphi \in \mathcal{S}$,
φ has a separating model implies $\mu(\varphi) > 0$.

- Dropping the 'separating' conflicts with the Gaifman condition.
- Cournot requires sentences, not interpretations,
  to have strictly positive probability,
  so is applicable even for uncountable model classes.

Black Ravens – Again

Let μ be Gaifman and Cournot, then:

$$\mu(\forall i.B(i)\,|\,B(1)\wedge ...\wedge B(n))$$

$$= \frac{\mu(\forall i.B(i))}{\mu(B(1)\wedge ...\wedge B(n))} \qquad \text{[Def. of } \mu(\varphi|\psi) \text{ and } \forall i.B(i)\to B(i)]$$

$$\overset{n\to\infty}{\longrightarrow} \frac{\mu(\forall i.B(i))}{\mu(\forall i.B(i))} \qquad [\mu \text{ is Gaifman}]$$

$$= 1 \qquad [\mu \text{ is Cournot}]$$

Eureka! Finally it works! This generalizes: Gaifman and Cournot are sufficient and necessary for confirming universal hypotheses.

Theorem (confirmation of universal hypotheses)

μ can confirm all universal hypotheses that have a separating model $\Leftrightarrow$ *μ is Gaifman and Cournot*

But: Do such μ exist? This is not the end but a start!

Constructing a Gaifman&Cournot Prior

Construction

- Enumerate the countable set of sentences with a separating model, $\chi_1, \chi_2, \ldots$
- For each sentence, χ_i, choose a separating interpretation that makes it true.
- Add probability mass $\frac{1}{i(i+1)}$ to that interpretation.
- Define μ^* to be the probability on this countable set of interpretations.
- Define μ to be the corresponding distribution over sentences.

Theorem (The μ constructed above is Gaifman and Cournot)

Minimum More Informative Probability

Given:

- a (Gaifman&Cournot) prior distribution μ over sentences, and
- a self-consistent set of constraints on probabilities: $\rho(\varphi_1) = a_1, ..., \rho(\varphi_n) = a_n$ given for *some* sentences $\varphi_1, ..., \varphi_n$.

Find:

- the distribution ρ that is minimally more informative than μ that meets the constraints. (KL-divergence)

Example

Given a prior distribution μ, adjust it so that it obeys the constraints:

A $\rho\,(\forall x.\forall y.x < 6 \Rightarrow y > 6) = 0.7$

B $\rho\,((\mathit{flies}\ \mathit{Tweety})) = 0.9$

C $\rho\,((\mathit{commutative}\ +)) = 0.9999$

Relative Entropy (KL)

Definition (relative entropy on sentences and interpretations)

Given any enumeration of all sentences $\varphi_1, \varphi_2, ...$, let

$$\psi_{n,S} := (\bigwedge_{i\in S} \varphi_i) \wedge (\bigwedge_{j\in\{1:n\}\setminus S} \neg\varphi_j) \quad \text{with} \quad S \subseteq \{1:n\}.$$

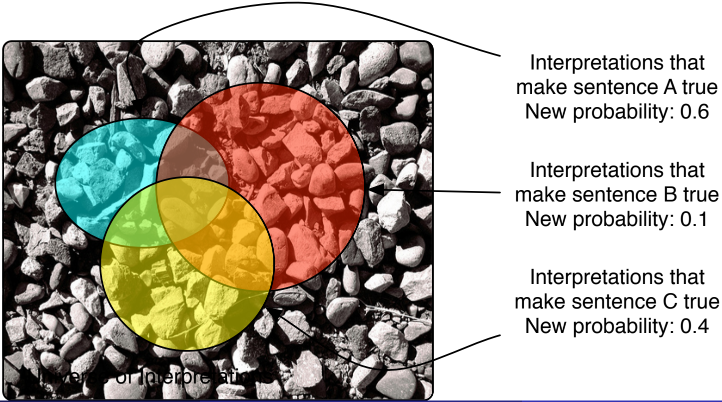
$$\text{Then} \quad \text{KL}(\rho||\mu) := \lim_{n\to\infty} \sum_{S\subseteq\{1:n\}} \rho(\psi_{n,S}) \log \frac{\rho(\psi_{n,S})}{\mu(\psi_{n,S})}$$

$$\text{KL}(\rho^*||\mu^*) := \int_{\mathcal{I}} \log \frac{d\rho^*}{d\mu^*}(I) d\rho^*(I)$$

Theorem ($\text{KL}(\rho||\mu) = \text{KL}(\rho^*||\mu^*)$)

Minimum Relative Entropy

- We have a set of interpretations, and a distribution, μ
- A set of constraints, which partition space of interpretations via $\psi_{n,S}$
- The distribution $\rho = \arg\min_\rho\{\mathrm{KL}(\rho||\mu) : \rho(\varphi_1) = a_1, ..., \rho(\varphi_n) = a_n\}$ that minimizes relative entropy KL is a multiplicative re-weighting, with constant weight across each partition:



Outlook

More in the paper

- Alternative tree construction (similar to ψ_S).
- General characterizations of probabilities that satisfy some or all of our criteria.
- Various (counter) examples of (strong) (non)Cournot and/or Gaifman probabilities and (non)separating interpretations.

More left for future generations [see www.hutter1.net/official/students.htm]

- Combine probability inside and outside sentences
- Incorporate ideas from Solomonoff induction to get optimal priors.
- Include description operator(s) (ι, ε).
- develop approximation schemes for the different currently incomputable aspects of the general theory.
- Develop a formal (incomplete, approximate) reasoning calculus

Summary: Our paper ...

- Shows that a function from sentences in a higher order logic to $\mathbb{R}$ gives a well defined probability distribution
- Extends two conditions for useful priors to the higher order setting
- Gives a theoretical construction for a prior that meets the conditions
- Gives general characterizations of probabilities that meet the conditions.
- Gives various (counter) examples of (strong) (non)Cournot and/or Gaifman probabilities and (non)separating interpretations.
- Notes that minimum relative entropy inference is well defined in this setting.

Achievement (unification of probability & logic & learning)

The results are a step towards a globally consistent and empirically satisfactory unification of probability and logic.

References

M. Hutter, K.S. Ng, J.W. Lloyd, and W.T.B. Uther.
Probability on Sentences in an Expressive Logic
Journal of Applied Logic, to appear, 2012.

H. Gaifman and M. Snir.
Probabilities over rich languages, testing and randomness.
The Journal of Symbolic Logic, 47(3):495–548, 1982.

W.M. Farmer.
The seven virtues of simple type theory.
Journal of Applied Logic, 6(3):267–286, 2008.

K.S. Ng, J.W. Lloyd, and W.T.B. Uther.
Probabilistic modelling, inference and learning using logical theories.
Annals of Mathematics and Artificial Intelligence, 54:159–205, 2008.

S. Rathmanner and M. Hutter.
A philosophical treatise of universal induction.
Entropy, 13(6):1076–1136, 2011.