Unifying Probability and Logic for Learning

Marcus Hutter
Research School of Computer Science
The Australian National University
marcus.hutter@anu.edu.au

John W. Lloyd
Research School of Computer Science
The Australian National University
john.lloyd@anu.edu.au

Kee Siong Ng
EMC Greenplum and
The Australian National University
keesiong.ng@emc.com

William T. B. Uther
National ICT Australia and
University of New South Wales
william.uther@nicta.com.au

Abstract

[1] Uncertain knowledge can be modeled by using graded probabilities rather than binary truth-values, but so far a completely satisfactory integration of logic and probability has been lacking. In particular the inability of confirming universal hypotheses has plagued most if not all systems so far. We address this problem head on. The main technical problem to be discussed is the following: Given a set of sentences, each having some probability of being true, what probability should be ascribed to other (query) sentences? A natural wish-list, among others, is that the probability distribution (i) is consistent with the knowledge base, (ii) allows for a consistent inference procedure and in particular (iii) reduces to deductive logic in the limit of probabilities being 0 and 1, (iv) allows (Bayesian) inductive reasoning and (v) learning in the limit and in particular (vi) allows confirmation of universally quantified hypotheses/sentences. We show that probabilities satisfying (i)-(vi) exist, and present necessary and sufficient conditions (Gaifman and Cournot). The theory is a step towards a globally consistent and empirically satisfactory unification of probability and logic.

Keywords

expressive languages; probability on sentences; Gaifman; Cournot; Bayes; induction; confirmation; learning; prior; knowledge; entropy.

"The study of probability functions defined over the sentences of a rich enough formal language yields interesting insights in more than one direction."

— Haim Gaifman (1982)

1 Introduction

Motivation. Sophisticated computer applications generally require expressive languages for knowledge representation and reasoning. In particular, such languages need to be able to represent both structured knowledge and uncertainty [Nilsson, 1986; Halpern, 2003; Muggleton, 1996; De Raedt and Kersting, 2003; Richardson and Domingos, 2006; Hájek, 2001; Williamson, 2002].

A key goal of this research is that of integrating logic and probability, a problem that has a history going back around 300 years and for which at least three main threads can be discerned: The oldest by far is the philosophical/mathematical thread that can be traced via Boole in 1854 back to Jacob Bernoulli in 1713. An extensive historical account of this thread can be found in [Hailperin, 1996]; the idea of putting probabilities on sentences goes back to before [Łos, 1955] which contains references to even earlier material; the important Gaifman condition appeared in [Gaifman, 1964] and was further developed in [Gaifman and Snir, 1982]; in [Scott and Krauss, 1966] the theory is developed for infinitary logic; overviews of more recent work from a philosophical perspective can be found in [Hájek, 2001; Williamson, 2002; 2008b]. The second thread is that of the knowledge representation and reasoning community in artificial intelligence, of which [Nilsson, 1986; Halpern, 1990; Fagin and Halpern, 1994; Halpern, 2003; Shirazi and Amir, 2007] are typical works. The third thread is that of the machine learning community in artificial intelligence, of which [Muggleton, 1996; De Raedt and Kersting, 2003; Richardson and Domingos, 2006; Milch and Russell, 2007; de Salvo Braz, 2007; Kersting and De Raedt, 2007; Pfeffer, 2007; Goodman *et al.*, 2008] are typical works. We admit that this categorization is rather terse, coarse, and incomplete.

An important and useful technical distinction that can be made between these various approaches is that the combination of logic and probability can be done externally or internally [Williamson, 2008b]: in the external view, probabilities are attached to sentences in some logic; in the internal view, sentences incorporate statements about probability. One can even mix the two cases so that probabilities appear both internally and externally [Halpern, 1990]. This paper takes the

[1] Hutter *et al.* [2013] contains all technical details and proofs and more discussion.

external view, leaving the combination with the internal view for future work.

Main aim. These considerations lead to the main technical issue studied in this paper:

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

We build on the work of Gaifman [1964] whose paper with Snir [1982] develops a quite comprehensive theory of probabilities on sentences in first-order Peano arithmetic. We take up these ideas, using non-dogmatic priors [Gaifman and Snir, 1982] and additionally the minimum relative entropy principle as in [Williamson, 2008a], but for general theories and in a higher-order setting. We concentrate on developing probabilities on sentences in a higher-order logic. This sets the stage for combining it with the probabilities inside sentences approach [Ng and Lloyd, 2009; Ng *et al.*, 2008].

Summary of key concepts. Section 3 gives the definition of probabilities on sentences (Definition 1) and shows their close connection with probabilities on interpretations. Gaifman [1964] (generalized in Definition 6) introduced a condition, called Gaifman in [Scott and Krauss, 1966], that connects probabilities of quantified sentences to limits of probabilities of finite conjunctions. In our case, it effectively restricts probabilities to separating interpretations while maintaining countable additivity.

While generally accepted in probability theory (Definition 2), some circles argue that countable additivity (CA) does not have a good philosophical justification, and/or that it is not needed since real experience is always finite, hence only non-asymptotic statements are of practical relevance, for which CA is not needed. On the other hand, it is usually much easier to first obtain asymptotic statements which requires CA, and then improve upon them. Furthermore we will show that CA can guide us in the right direction to find good finitary prior probabilities.

Another principle which has received much less attention than CA but is equally if not more important is that of Cournot [Cournot, 1843; Shafer, 19 May 2006]: An event of probability (close to) zero singled out in advance is physically impossible; or conversely, an event of probability 1 will physically happen for sure. In short: zero probability means impossibility. The history of the semantics of probability is stony [Fine, 1973]. Cournot's "forgotten" principle is one way of giving meaning to probabilistic statements like, "the relative frequency of heads of a fair coin converges to 1/2 *with probability 1*". The contraposition of Cournot is that one must assign non-zero probability to possible events. If "events" are described by sentences and "possible" means it is possible to satisfy these sentences, i.e. they possess a model, then we arrive at the strong Cournot principle that satisfiable sentences should be assigned non-zero probability. This condition has been appropriately called 'non-dogmatic' in [Gaifman and Snir, 1982]. As long as something is not proven false, there is a (small) chance it is true in the intended interpretation. This non-dogmatism is crucial in Bayesian inductive reasoning, since no evidence (however strong) can increase a zero prior belief to a non-zero posterior belief [Rathmanner and Hutter, 2011]. The Gaifman condition is inconsistent with the strong Cournot principle, but consistent with a weaker version (Definition 8). Probabilities that are Gaifman and (plain, not strong) Cournot allow learning in the limit (Theorem 9 and Corollary 11).

A standard way to construct (general / Cournot / Gaifman) probabilities on sentences is to construct (general / non-dogmatic / separating) probabilities on interpretations, and then transfer them to sentences (Proposition 4). At the same time we give model-theoretic characterizations of the Gaifman condition (Theorem 7). We also give a particularly simple construction of a probability that is Cournot and Gaifman (Theorem 10) and a complete characterization of general/Cournot/Gaifman probabilities in [Hutter *et al.*, 2013].

At the end of Section 3 we briefly and in [Hutter *et al.*, 2013] we fully consider the important practical situation of whether and how a real-valued function on a set of sentences can be extended to a probability on all sentences; a method for determining such probabilities is given. Prior knowledge and data constrain our (belief) probabilities in various ways, which we need to take into account when constructing probabilities. Prior knowledge is usually given in the form of probabilities on sentences like "the coin has head probability 1/2", or facts like "all electrons have the same charge", or non-logical axioms like "there are infinitely many natural numbers". They correspond to requiring their probability to be 1/2, extremely close to 1, and 1, respectively. It is therefore necessary to be able to go from probabilities on sentences to probability on interpretations (Proposition 3). Seldom does knowledge constrain the probability on all sentences to be uniquely determined. In this case it is natural to choose a probability that is least dogmatic or biased [Nilsson, 1986; Williamson, 2008a]. The minimum relative entropy principle can be used to construct such a unique minimally more informative probability that is consistent with our prior knowledge.

Section 4 outlines how the developed theory might be used and approximated in autonomous reasoning agents. In particular, certain knowledge, learning in the limit (Corollary 11) and the infamous black raven paradox are discussed. Section 5 contains a brief summary and future research directions.

We start with some preliminaries in the following Section 2.

2 Preliminaries

This section sets the stage for the subsequent theoretical development and applications. We introduce the black raven hypothesis, used as a running example to illustrate and motivate the theory. Then we state a natural wish-list for the prior probability distribution, and the technical requirements they translate into. This also allows us to describe the intuition behind our main results, before delving into technicalities in Section 3. Finally the used logic is outlined.

Induction example: black ravens. As discussed, the main goal of this paper is to unify probability and logic for learn-

ing. We illustrate and motivate the theory developed in this section by a running example, namely the confirmation of universal hypotheses. The black raven hypothesis is an infamous instantiation [Earman, 1993; Maher, 2004]. It is technically very simple, while still most reasoning systems fail on it.

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", that is $\forall x.B(x)$. Intuitively, observing more and more black ravens with no counter-examples increases our confidence in the hypothesis. So a plausible requirement on any inductive reasoning system is that Probability$(\forall x.B(x) \mid B(1) \wedge ... \wedge B(n))$ tends to 1 for $n \to \infty$.

Real-world problems are much more complex, but most reasoning systems fail already on this apparently simple example. For instance, Bayes/Laplace rule and Carnap's confirmation theory fail, but Solomonoff induction works [Rathmanner and Hutter, 2011]. A more complex example is given in Section 4. Finally note that the (full) black raven paradox is more complicated and will not be discussed here.

Wish-list. Expressive logic languages are ideally suited for representing and reasoning about structured knowledge. Uncertain knowledge can be modeled by assigning graded probabilities rather than binary truth-values to sentences. Together this suggests to put probabilities on sentences. As stated in the introduction, the main technical problem considered is: Given a set of sentences, each having some probability of being true, what probability should be ascribed to other (query) sentences? This sets the stage for combining it with the probabilities inside sentences approach [Ng and Lloyd, 2009; Ng *et al.*, 2008].

A natural wish List (among others) is that 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. We will see that this wish-list translates into the following technical requirements for a prior probability: It needs to be

(P) consistent with the standard axioms of **P**robability,
(CA) including **C**ountable **A**dditivity,
(C) non-dogmatic $\widehat{=}$ **C**ournot
$\widehat{=}$ zero probability means impossibility
$\widehat{=}$ whatever is not provably false is assigned probability larger than 0.
(G) separating $\widehat{=}$ **G**aifman
$\widehat{=}$ existence is always witnessed by terms
$\widehat{=}$ logical quantifiers over variables can be replaced by meta-logical quantification over terms.

Main results. In the next section we will give suitable formalizations of all requirements. We give one explicit "construction" of such probabilities. Proofs that they satisfy all our criteria, general characterizations of probabilities that satisfy some or all of the criteria, and various (counter) examples of (strong) (non)Cournot and/or Gaifman probabilities and (non)separating interpretations can be found in [Hutter *et al.*, 2013].

We also give 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 Cournot and Gaifman prior (1) exists, and is (2) consistent with our prior knowledge, (3) minimally more informative, (4) unique, and (5) suitable for inductive inference. Section 4 outlines how to use and approximate the theory for autonomous reasoning agents.

On the choice of logic. In practice, ignoring computational considerations, the more expressive the logic the better. Higher-order logic, also called simple type theory (STT), is such an expressive logic. In Hutter *et al.* [2013] we fully develop the theory for STT with Henkin semantics without description operator for countable alphabet.

The major ideas though work in many logics (e.g. first order), but there are important and subtle pitfalls to be avoided. Due to limited space, we will here abstract away from and gloss over the details of the used logic.

As usual we have boolean operations $\top, \bot, \wedge, \vee, \to$, quantifiers $\forall x, \exists y$, closed terms t, sentences φ, χ, formula $\psi(x)$ with a single free variable x, universal hypothesis/sentence $\forall x.\psi(x)$, usually equality $=$, and in STT abstraction λz, but this is not needed here.

3 Theory

We define probabilities μ over sentences φ in the usual way [Halpern, 1990]: $\mu(\varphi)$ is the probability that φ is true in the intended interpretation, or $\mu(\varphi)$ is the subjective probability held by an agent that sentence φ holds in the real world. It should satisfy the basic axioms of probability, and hence has the usual properties. Only countable Additivity (CA) enters later and differently, since finitary logics lack infinite conjunctions of sentences.

Definition 1 (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:* $\mu(\varphi|\chi) := \mu(\varphi \wedge \chi)/\mu(\chi)$.

A sentence φ is said to be valid, if it is true in all (Henkin) interpretations. We also define probabilities on interpretations, which is closer to conventional measure theory. Let

mod(φ) be the class of (Henkin) interpretations in which φ is true, and $\mathcal{I}$:= *mod*($\top$) be the class of all (Henkin) interpretations, and $\mathcal{B}$ be the σ-algebra generated by $\{$*mod*$(\varphi) : \varphi \in \mathcal{S}\}$. Then:

Definition 2 (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 it holds that $\mu^*(\bigcup_{i\in I} A_i) = \sum_{i\in I} \mu^*(A_i)$.*

Probability on sentences $\Leftrightarrow$ interpretations. There is a close relationship between probabilities on sentences and probabilities on interpretations. This allows us to exploit (some) results from measure theory, valid for the latter, also for the former.

Proposition 3 ($\mu \Rightarrow \mu^*$) *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 $\varphi \in \mathcal{S}$.*

The proof uses compactness of the class of (Henkin) interpretations $\mathcal{I}$ and Caratheodory's unique-extension theorem. The converse is elementary:

Proposition 4 ($\mu^* \Rightarrow \mu$) *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 $\mathcal{S}$.*

Problems. Consider the black raven example: Intuitively, knowledge of $\{B(1), B(2), ...\} \equiv \{B(i) : i \in \mathbb{N}\}$ should imply $\forall x.B(x)$.

Problem is that this is not true in all models. There are non-standard models of the natural numbers in which $x = n$ is invalid for all $n = 1, 2, 3, ...$ The reason is that the natural numbers have neither a categorical axiomatization in first order logic, nor in STT with Henkin semantic. They *do* in STT with normal semantics, but there compactness and hence the crucial Proposition 3 fails. So in either case we have a problem.

The solution is to exclude such unwanted interpretations. The natural generalization of "$1, 2, 3, ...$" for general theories is "all terms t".

Definition 5 (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 $\psi\{x/t\}$, where $\psi\{x/t\}$ is ψ with all free x replaced by t.*

Informally this means that existence is always witnessed by terms. For objects to exist we must be able to name them. It is important to note that our vocabulary from which the closed terms are constructed is fixed up front and the same for all I. Otherwise we could trivially make every interpretation separating by adding sufficiently many new constants to the theory, as e.g. done in Henkin's construction. We need to avoid such new constants since they would ruin induction. In complete analogy to above, let $\widehat{mod}(\varphi)$ be the set of separating models of φ, $\widehat{\mathcal{I}} = \widehat{mod}(\top)$ be the set of all separating interpretations, and $\widehat{\mathcal{B}}$ be the σ-algebra generated by $\{\widehat{mod}(\varphi) : \varphi \in \mathcal{S}\}$. Note that all $\widehat{mod}(\varphi)$ are $\mathcal{B}$-measurable.

Next we effectively avoid non-separating interpretations by requiring the probability on them to be zero:

Definition 6 (Gaifman condition) *We call μ Gaifman iff*

$$\mu(\forall x.\psi(x)) = \lim_{n\to\infty} \mu(\textstyle\bigwedge_{i=1}^{n} \psi\{x/t_i\})$$

for all ψ, where $t_1, t_2, ...$ is an enumeration of (representatives of) all closed terms (of same type as x).

Informally this means that logical quantifiers over variables can be replaced by meta-logical quantification over terms: With 'representative' we mean that one term per $=$-equivalence class is sufficient. For the theory of natural numbers, all terms (of type *Nat*), equal $\underline{1}$ or $\underline{2}$ or ..., e.g. $t = \underline{5} + \underline{3}$ equals $\underline{8}$, hence does not need to be listed separately.

Theorem 7 ($\mu^*(\mathcal{I} \setminus \widehat{\mathcal{I}}) = 0 \;\Leftrightarrow\; \mu$ is Gaifman) *For any probability $\mu : \mathcal{S} \to \mathbb{R}$ on sentences and probability $\mu^* : \mathcal{B} \to \mathbb{R}$ on interpretations (one-to-one) related by $\mu^*($mod$(\varphi)) = \mu(\varphi)$ it holds that: $\mu^*(\mathcal{I} \setminus \widehat{\mathcal{I}}) = 0 \Leftrightarrow \mu$ Gaifman.*

Induction still does not work. Unfortunately, even μ satisfying the Gaifman condition may fail to confirm universal hypotheses. The reason is that $\mu(\forall x.B(x) \mid B(1)\wedge ...\wedge B(n)) \equiv 0$ if $\mu(\forall x.B(x)) = 0$. This is the infamous Zero-Prior problem in philosophy of induction. If your prior excludes some hypothesis, no amount of evidence can confirm it. Carnap's and most other confirmation theories fail, since they (implicitly & unintentionally) have $\mu(\forall x.B(x)) = 0$. Why is this problem hard? "Naturally" $\mu(\forall x.B(x)) \leq \mu(B(1) \wedge ... \wedge B(n)) \to 0$. Think of independent events with probability $p < 1$, then $p \cdot p \cdot p \cdots \to 0$. But it's not hopeless: We just *demand* $\mu(\forall x.\psi(x)) > 0$ for all ψ for which this is possible and/or reasonable, which turns out to be the φ that have separating models.

We call this Cournot's principle: Informally stated, probability zero/one means impossibility/certainty, or whatever is not provably false is assigned probability larger than 0, or all (sensible) prior probabilities should be non-zero, or be as non-dogmatic as possible. Formally:

Definition 8 (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 $\mu(\varphi) > 0$.*

We cannot drop the 'separating', since this would then conflict with the Gaifman condition. Note that Cournot requires sentences, not interpretations, to have strictly positive probability, so is applicable even for uncountable model classes.

Black ravens – again. Consider a theory in which all terms (of type *Nat*) represent natural numbers. Let μ be Cournot and Gaifman, then:

$$\begin{aligned}
&\mu(\forall x.B(x) \mid B(1) \wedge ... \wedge B(n)) \\
&\quad = \frac{\mu(\forall x.B(x))}{\mu(B(1) \wedge ... \wedge B(n))} && \left[\begin{array}{l}\text{Def. of } \mu(\varphi|\psi) \text{ and} \\ \forall x.B(x) \to B(i)\end{array}\right] \\
&\quad \xrightarrow{n\to\infty} \frac{\mu(\forall x.B(x))}{\mu(\forall x.B(x))} && [\mu \text{ is Gaifman}] \\
&\quad = 1 && [\mu \text{ is Cournot}]
\end{aligned}$$

Finally induction works! This example generalizes: The Cournot and Gaifman conditions are sufficient and necessary for confirming universal hypotheses.

Theorem 9 (confirmation of universal hypotheses) *μ can confirm all universal hypotheses that have a separating model $\Leftrightarrow$ μ is Cournot and Gaifman.*

What remains to be shown is whether such μ actually exist. General characterizations are given in [Hutter *et al.*, 2013]. A particularly simple "construction" is as follows:

Theorem 10 (Constructing a Cournot and Gaifman prior μ) *The following μ is Cournot and Gaifman:*

- *Enumerate the countable set of sentences that have a separating model, $\chi_1, \chi_2, \ldots$*
- *For each sentence, χ_i, choose a separating interpretation that makes it true.*
- *Assign 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.*

Alternatively one can enumerate *all* sentences φ_1, φ_2, φ_3, ..., and in an infinite binary tree label each left (right) branch at depth n with $\neg\varphi_n$ (φ_n) and assign probabilities to each node as detailed in [Hutter *et al.*, 2013, Thm.52], which in turn defines μ.

Very powerful Cournot and Gaifman (C&G) probabilities can be constructed as follows: Let $\mathcal{M} = \{\nu_1, \nu_2, ...\}$ be any finite or countable class of Gaifman probabilities of interest. These are usually priors that are potentially true, e.g. i.i.d. probabilities such as $\nu(B(1) \wedge \cdots \wedge B(n)) = (\frac{1}{2})^n$ which is not Cournot. Now define the mixture $\xi(\varphi) := \sum_{\nu_i \in \mathcal{M}} \frac{\nu_i(\varphi)}{i(i+1)}$, which is also Gaifman and mimics Solomonoff's construction [Solomonoff, 1964; Hutter, 2005]. If for every sentence χ that has a separating model there exists a $\nu \in \mathcal{M}$ such that $\nu(\chi) > 0$, then ξ is also Cournot, since $\xi(\chi) > \nu(\chi) > 0$. If this is not the case, we can simply add some/one/any C&G prior μ to $\mathcal{M}$, e.g. the one from Theorem 10, which makes ξ C&G. Since ξ dominates all $\nu \in \mathcal{M}$, the Merging-of-Opinions theorem [Blackwell and Dubins, 1962] guarantees that ξ converges to ν in total variation with ν probability 1 for any $\nu \in \mathcal{M}$. This means, while the Cournot condition rules out e.g. i.i.d. distributions, there are C&G probabilities ξ that converge to them, provided the data warrant it, and that is usually all we need.

While asymptotic convergence works equally for any C&G probability, the degree of confirmation from finite sample size depends on the specific construction. To achieve fast convergence for μ constructed in Theorem 10 one should sort sentences in decreasing order of "relevance" and pick "natural" models. note that $1/i(i+1)$ is nearly as uniform as possible, hence the order dependence is benign compared to e.g. 2^{-i}.

Minimum more informative probability. Knowledge is usually given as constraints on some probability distribution ρ. Hard facts have $\rho(\text{fact}) = 1$, while uncertain knowledge has $0 < \rho < 1$. This still leaves many choices for ρ. In our context it is natural to start with some C&G prior μ, and find a "minimally more informative" ξ consistent with the knowledge base. A natural notion of "minimally more informative" is the minimum relative entropy.

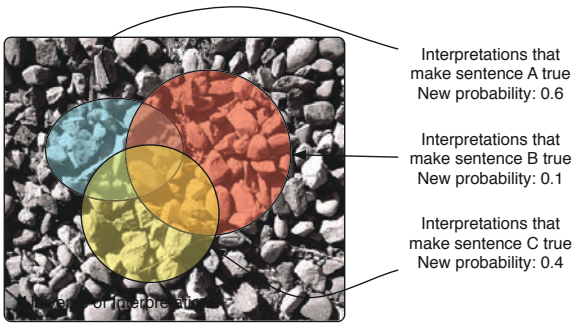
More formally, the task is: Given a C&G prior distribution ξ over sentences, and a self-consistent set of constraints on probabilities, $\rho(\varphi_1) = a_1, ..., \rho(\varphi_n) = a_n$ given for *some* sentences $\varphi_1, ..., \varphi_n$. Find the distribution ρ that minimizes $\mathrm{KL}(\rho||\xi)$ under the constraints.

For example, given a prior distribution ξ, minimally 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) $\rho\,((\mathit{commutative}\ +)) = 0.9999$

The solution consists of the following steps: (i) choose a prior ξ, e.g. the one in Theorem 10. (ii) determine the consistency of the knowledge base $\{\rho(\varphi_i) = a_i\}$. Sufficient conditions are given in [Hutter *et al.*, 2013]. (iii) $\mathrm{KL}(\rho||\xi)$ can be defined as $\mathrm{KL}(\rho^*||\xi^*)$, where the latter is the standard measure-theoretic definition. We have derived explicit finite expression of $\mathrm{KL}(\rho||\xi)$ without reference to probabilities on interpretations, and finite equation systems for minimizing $\mathrm{KL}(\rho||\xi)$ w.r.t. ρ under constraints $\{\rho(\varphi_i) = a_i\}$.

In effect, the constraints partition the space of (separable) interpretations $\widehat{\mathcal{I}}$, and the ρ^* corresponding to the distribution $\rho = \arg\min_\rho\{\mathrm{KL}(\rho||\xi) : \rho(\varphi_1) = a_1, ..., \rho(\varphi_n) = a_n\}$ that minimizes the relative entropy KL is a multiplicative reweighting of ξ, with constant weight across each partition. This is depicted in the example below, where pixels correspond to interpretations, their intensity to their probability, and each (mixed) color to a region with uniform multiplicative reweight. All derivations and equations can be found in [Hutter *et al.*, 2013].



4 User Manual

This section outlines how (approximations of) the theory developed in Section 3 might be used in autonomous reasoning agents. We discuss the special case of certain knowledge and how it can be used to make inferences about statements that are not logical implications of the knowledge base. For instance, if our agent has observed a large number of ravens which are all black without exception, how strongly should it belief in the hypothesis that "all ravens are black"? We

construct an agent that can learn in the limit in the usual time-series forecasting setting with an observation sequence indexed by natural numbers.

Certain knowledge. A common case of knowledge is a set of sentences φ_i, each having degree of belief 1 (that is, $\mu_0(\varphi_i) = 1$, for $i = 1, \ldots, n$). In other words, there is certainty that each φ_i is valid in the intended interpretation. This corresponds to non-logical axioms in a theory. Let ξ be a Cournot probability and suppose that μ is minimally more informative than ξ given μ_0. For this situation, one can show that μ satisfies

$$\mu(\varphi) = \xi(\varphi \,|\, \varphi_1 \wedge \cdots \wedge \varphi_n), \tag{1}$$

for $\varphi \in \mathcal{S}$. Consequently, either $\varphi_1 \wedge \cdots \wedge \varphi_n$ is satisfiable (leading directly to the above definition for μ) or else it is not, in which case there are no solutions and μ cannot be defined at all.

A further special case beyond the one just considered is when φ is a logical consequence of $\varphi_1 \wedge \cdots \wedge \varphi_n$. In this case,

$$\mu(\varphi) \;=\; \xi(\varphi \,|\, \varphi_1 \wedge \cdots \wedge \varphi_n) \;=\; \frac{\xi(\varphi_1 \wedge \cdots \wedge \varphi_n)}{\xi(\varphi_1 \wedge \cdots \wedge \varphi_n)} \;=\; 1,$$

as one would expect. Similarly when $\neg\varphi$ is logical consequence, then $\mu(\varphi) = 0$.

Note that, while it is important that the prior ξ be Cournot, it is just as important that the posterior μ be allowed not to be Cournot. The prior should be Cournot so that the KL divergence is as widely defined as possible or, more intuitively, to make sure sentences having a separating model are *not forced* to have μ-probability 0. On the other hand, the probability μ should be *allowed* to be 0 on sentences having a separating model since the evidence in the form of the probabilities on $\varphi_1, \ldots, \varphi_n$ may imply this. This is apparent, for example, for the case where each φ_i has probability 1: according to this evidence, any sentence (even one having a separating model) that is disjoint from $\varphi_1 \wedge \cdots \wedge \varphi_n$ must have μ-probability 0.

Black ravens. Let the evidence consist of the sentences $B(1), \ldots, B(n)$, whence $\varphi_i \equiv B(i)$, for $i = 1, \ldots, n$. Let $\mu_0 : \{B(1), \ldots, B(n)\} \to [0, 1]$ be defined by $\mu_0(B(i)) = 1$, for $i = 1, \ldots, n$. Thus the degree of belief that the ith raven is black is 1, for $i = 1, \ldots, n$. Suppose that ξ is an uninformative prior that is C&G. Since a-priori there are no constraints (on B), this implies that $\xi(\forall x.B(x)) > 0$. Let μ be a probability that is minimally more informative than ξ given μ_0. Thus μ is given by (1).

Now consider the sentence $\forall x.B(x)$. This is clearly not a logical consequence of the evidence, but one can use μ to ascribe a degree of belief that it is true and, furthermore, investigate what happens to this probability as the number of black ravens increases. Equation (1) and $\mu_0(B(i)) = 1$, for $i = 1, \ldots, n$, and then Theorem 9 applied to C&G ξ show that

$$\mu(\forall x.B(x)) \;=\; \xi(\forall x.B(x) \,|\, B(1) \wedge \cdots \wedge B(n)) \;\stackrel{n\to\infty}{\longrightarrow}\; 1$$

Thus, as the number of observed black ravens increases, the degree of belief that all ravens are black approaches 1. Of course this also implies the weaker statement that our belief in the next raven being black tends to one:

$$\xi(B(n{+}1) \,|\, B(1) \wedge \cdots \wedge B(n)) \quad \stackrel{n\to\infty}{\longrightarrow} \quad 1$$

Naive black ravens. Continuing the preceding example, suppose given the evidence $B(1), \ldots, B(n)$, each having probability 1, one wants to know the degree of belief for $B(n{+}1)$. Most probabilistic reasoning systems, if they have at all the ability to provide prior distributions, give $\xi(B(1) \wedge \cdots \wedge B(n)) = (\frac{1}{2})^n$ or similar, which can usually be traced back to a (naive) application of the maximum entropy or indifference principle, and/or to first assigning probabilities to quantifier-free probabilities and then extending them to quantified formulas. In this case

$$\begin{aligned}&\xi(B(n{+}1) \,|\, B(1) \wedge \cdots \wedge B(n)) \\ &= \frac{\xi(B(1) \wedge \cdots \wedge B(n) \wedge B(n{+}1))}{\xi(B(1) \wedge \cdots \wedge B(n))} \;=\; \frac{1}{2}.\end{aligned}$$

Thus, for this prior, knowing the evidence so far, even for large n, does not give any information about $B(n{+}1)$. But it gets worse: Assume ξ is somehow extended to a probability on all $\mathcal{S}$. Then for any $m \geq n$,

$$\begin{aligned}&\xi(\forall x.B(x) \,|\, B(1) \wedge \cdots \wedge B(n)) \\ &\leq \xi(B(1) \wedge \cdots \wedge B(m) \,|\, B(1) \wedge \cdots \wedge B(n)) \;=\; (\tfrac{1}{2})^{m-n}\end{aligned}$$

hence $\xi(\forall x.B(x) \,|\, B(1) \wedge \cdots \wedge B(n)) \equiv 0$ for all n, i.e. universal hypotheses can not be confirmed. Even more seriously, we would be absolutely sure that non-black ravens exist

$$\xi(\exists i.\neg B(i) \,|\, B(1) \wedge \cdots \wedge B(n)) \equiv 1$$

and no number of observed black ravens n without any counter examples will ever convince us otherwise. The crucial requirement to avoid these problems was to include quantified sentences when constructing a prior and ensure it is Cournot (even when only making inferences about unquantified sentences like $B(n{+}1)$).

Corollary 11 (learning in the limit) *Let ψ be a formula with free variable x of type Nat, μ be a Gaifman probability on sentences, and $\mu(\forall x.\psi(x)) > 0$. Then*

$$\lim_{n\to\infty} \mu(\forall x.\psi(x) \,|\, \psi(\underline{0}) \wedge \cdots \wedge \psi(\underline{n})) \;=\; 1$$

This generalizes the black raven example and follows from Theorem 9. In particular, learning in the limit is possible for the C&G probability constructed in Theorem 10, provided $\forall x.\psi(x)$ has a separating model.

The proof crucially exploits that $\underline{0}, \underline{1}, \underline{2}, \ldots$ are representatives of all terms of type *Nat*. As discussed in [Hutter *et al.*, 2013], this would no longer be true had we introduced a description operator into our logic. Corollary 11 would break down and universal hypotheses over the natural numbers could not be inductively confirmed, not even asymptotically.

Approximations. The construction of C&G μ in Theorem 10 required to determine particular separating models for χ_i and

to determine whether they are also models of other sentences φ.

Assume we had some calculus to determining whether sentences have (no) separating model. Even an asymptotic or approximate or incomplete calculus may be of use. Fix a sequence on-the-fly of all sentences $\varphi_1, \varphi_2, \varphi_3, ...$ (once and for all). Determine the subsequence of all sentences $\chi_1 = \varphi_{j_1}, \chi_2 = \varphi_{j_2}, ...$ with separating models (on the fly).

In order to determine μ to accuracy $\varepsilon > 0$ for some finite number of sentences $\{\varphi_{i_1}, ..., \varphi_{i_n}\}$ of interest, we have to assign probability $\frac{1}{i(i+1)}$ "only" to χ_i for $i \leq m := \max\{\frac{1}{\varepsilon}, i_1, ..., i_n\}$, i.e. determine finitely many cases. If a new sentence $\varphi_{i_{n+1}}$ of interest "arrives" or higher precision is needed, m can be increased appropriately (that's what was meant with on-the-fly).

Work flow example for a simple inductive reasoning agent. Below we present an example of a fictitious inductive reasoning agent. It is fictitious, since many operations are incomputable. In practice one needs to employ approximations at various steps. How to do this is an open problem.

1. Assume the agent has been endowed with some background knowledge e.g. about kinetics, colors, biology, birds, etc. Its knowledge is represented in the form of a finite set of sentences $\{\varphi_1, \ldots, \varphi_n\}$ that hold for sure ($\mu_0(\varphi_i) = 1$ for some i) or with some probability $0 < \mu_0(\varphi_i) < 1$ for the other i.

2. In [Hutter *et al.*, 2013] we derive sufficient conditions (hierarchical, sub-additive, eligible) for μ_0 to be consistent. This task is akin to the general problem of maintaining consistent knowledge bases.

3. Next, use an approximation of a C&G ξ prior, e.g. as defined in Theorems 10 or the mentioned tree constructions as outlined above and detailed in [Hutter *et al.*, 2013]. The agent now constructs the minimally more informative probability μ, which has been shown to exist and be Gaifman.

4. Let $o_1, o_2, o_3, ...$ be the agent's life-time sequence of past and future observations of all kinds of objects, ravens and otherwise, all it has/will ever observe, e.g. o_n is what the agent sees n seconds after it has been switched on.

5. Assume current time is n, and the agent needs to hypothesize about the world to decide its next action, e.g. whether some observed regularity is "real". For instance, "if observation at time k is a raven, is it also black?". We can formalize this with a monadic predicate ψ for type *Nat* with the intended interpretation of $\psi(\underline{k})$ as "if observation at time k is a raven, it is black".

6. Of course the answer to $\psi(\underline{1}), ..., \psi(\underline{n})$ is immediate, since $o_1, ..., o_n$ have already been observed. If they are all true, the agent may start to wonder whether "all ravens are black", or formally, whether $\forall x.\psi(x)$ is true. Note that non-raven observations in the sequence are allowed.

7. If the agent is equipped with our inductive reasoning system, its degree of belief in this hypothesis is $\mu(\forall x.\psi(x)|\psi(\underline{1}) \wedge \cdots \wedge \psi(\underline{n}))$.

8. This result can be the basis for some decision process maximizing some utilities resulting in an informed action.

Is the degree of belief derived in Step 7 and used in Step 8 reasonable? At least asymptotically Corollary 11 ensures that in the limit the agent's belief tends to 1, which is very reasonable. So our system of inductive reasoning at least passes this test. Most other inductive reasoning systems have difficulties in getting this right [Rathmanner and Hutter, 2011].

5 Conclusion

This paper provided much of the foundation for the design of an integrated probabilistic reasoning system that can handle probabilities outside sentences.

We have shown that a function from sentences to $\mathbb{R}$ that is a well defined probability distribution with all of our criteria exists. In particular we gave a theoretical construction for a prior that meets the conditions, and showed that minimum relative entropy inference is well defined in this setting.

Besides proofs and more details and discussion, Hutter *et al.* [2013] additionally give general characterizations of probabilities that meet some or all of our criteria, and give various (counter) examples of (strong) (non)Cournot and/or Gaifman probabilities and (non)separating interpretations,

Overall, the results are a step towards a globally consistent and empirically satisfactory unification of probability and logic for learning.

There is much left for future research: To combine probabilities inside and outside sentences as in [Halpern, 1990], to incorporate ideas from Solomonoff induction to get optimal priors [Rathmanner and Hutter, 2011], to include the description operator(s) (ι, ε), and to investigate a number of other theoretical questions. The main challenge for the future lies in the discovery of reasonable approximation schemes for the different currently incomputable aspects of the general theory.

Acknowledgements. We thank the reviewers for their feedback. The research was partly supported by the Australian Research Council Discovery Project DP0877635 "Foundations and Architectures for Agent Systems". NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

References

[Blackwell and Dubins, 1962] D. Blackwell and L. Dubins. Merging of opinions with increasing information. *Annals of Mathematical Statistics*, 33:882–887, 1962.

[Cournot, 1843] A. A. Cournot. *Exposition de la théorie des chances et des probabilités*. L. Hachette, Paris, 1843.

[De Raedt and Kersting, 2003] L. De Raedt and K. Kersting. Probabilistic logic learning. *SIGKDD Explorations*, 5(1):31–48, 2003.

[de Salvo Braz, 2007] R. de Salvo Braz. *Lifted First-Order Probabilistic Inference*. PhD thesis, University of Illinois at Urbana-Champaign, 2007.

[Earman, 1993] J. Earman. *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory*. MIT Press, Cambridge, MA, 1993.

[Fagin and Halpern, 1994] R. Fagin and J.Y. Halpern. Reasoning about knowledge and probability. *Journal of the ACM*, 41(2):340–367, 1994.

[Fine, 1973] T. L. Fine. *Theories of Probability*. Academic Press, New York, 1973.

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

[Gaifman, 1964] H. Gaifman. Concerning measures in first order calculi. *Israel Journal of Mathematics*, 2(1):1–18, 1964.

[Goodman *et al.*, 2008] N. D. Goodman, V. K. Mansighka, D. Roy, K. Bonawitz, and J. B. Tenenbaum. Church: a language for generative models. In *Uncertainty in Artificial Intelligence*, 2008.

[Hailperin, 1996] T. Hailperin. *Sentential Probability Logic*. Lehigh University Press, 1996.

[Hájek, 2001] A. Hájek. Probability, logic and probability logic. In L. Goble, editor, *The Blackwell Guide to Philosophical Logic*, chapter 16, pages 362–384. Blackwell, 2001.

[Halpern, 1990] J.Y. Halpern. An analysis of first-order logics of probability. *Artificial Intelligence*, 46(3):311–350, 1990.

[Halpern, 2003] J.Y. Halpern. *Reasoning about Uncertainty*. MIT Press, 2003.

[Hutter *et al.*, 2013] M. Hutter, J.W. Lloyd, K.S. Ng, and W.T.B. Uther. Probabilities on sentences in an expressive logic. *Journal of Applied Logic*, 2013. in press.

[Hutter, 2005] M. Hutter. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*. Springer, Berlin, 2005.

[Kersting and De Raedt, 2007] K. Kersting and L. De Raedt. Bayesian logic programming: Theory and tool. In L. Getoor and B. Taskar, editors, *Introduction to Statistical Relational Learning*. MIT Press, 2007.

[Łos, 1955] J. Łos. On the axiomatic treatment of probability. *Colloquium Mathematicum*, 3:125–137, 1955.

[Maher, 2004] P. Maher. Probability captures the logic of scientific confirmation. In C. Hitchcock, editor, *Contemporary Debates in Philosophy of Science*, chapter 3, pages 69–93. Blackwell Publishing, 2004.

[Milch and Russell, 2007] B. Milch and S. Russell. First-order probabilistic languages: Into the unknown. In S. Muggleton, R. Otero, and A. Tamaddoni-Nezhad, editors, *Inductive Logic Programming: 16th International Conference, ILP 2006*, pages 10–24. Springer, LNAI 4455, 2007.

[Muggleton, 1996] S. Muggleton. Stochastic logic programs. In L. De Raedt, editor, *Advances in Inductive Logic Programming*, pages 254–264. IOS Press, 1996.

[Ng and Lloyd, 2009] K.S. Ng and J. W. Lloyd. Probabilistic reasoning in a classical logic. *Journal of Applied Logic*, 7(2):218–238, 2009. DOI:10.1016/j.jal.2007.11.008.

[Ng *et al.*, 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. DOI:10.1007/s10472-009-9136-7.

[Nilsson, 1986] N.J. Nilsson. Probabilistic logic. *Artificial Intelligence*, 28(1):71–88, 1986.

[Pfeffer, 2007] A. Pfeffer. The design and implementation of IBAL: A general-purpose probabilistic language. In Lise Getoor and Ben Taskar, editors, *Introduction to Statistical Relational Learning*, chapter 14. MIT Press, 2007.

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

[Richardson and Domingos, 2006] M. Richardson and P. Domingos. Markov logic networks. *Machine Learning*, 62:107–136, 2006.

[Scott and Krauss, 1966] D. Scott and P. Krauss. Assigning probabilities to logical formula. In J. Hintikka and P. Suppes, editors, *Aspects of Inductive Logic*, pages 219–264. North-Holland, 1966.

[Shafer, 19 May 2006] G. Shafer. Why did Cournot's principle disappear?, 19 May 2006. Presentation. Ecole des Hautes Etudes en Sciences Sociales, Paris. Slides, URL: http://www.glennshafer.com/assets/downloads/disappear.pdf.

[Shirazi and Amir, 2007] A. Shirazi and E. Amir. Probabilistic modal logic. In R.C. Holte and A. Howe, editors, *Proceedings of the 22nd AAAI Conference on Artificial Intelligence*, pages 489–495, 2007.

[Solomonoff, 1964] R. J. Solomonoff. A formal theory of inductive inference: Parts 1 and 2. *Information and Control*, 7:1–22 and 224–254, 1964.

[Williamson, 2002] J. Williamson. Probability logic. In D. Gabbay, R. Johnson, H.J. Ohlbach, and J. Woods, editors, *Handbook of the Logic of Inference and Argument: The Turn Toward the Practical*, volume 1 of *Studies in Logic and Practical Reasoning*, pages 397–424. Elsevier, 2002.

[Williamson, 2008a] J. Williamson. Objective bayesian probabilistic logic. *Journal of Algorithms*, 63(4):167–183, 2008.

[Williamson, 2008b] J. Williamson. Philosophies of probability. In A. Irvine, editor, *Handbook of the Philosophy of Mathematics, Volume 4 of the Handbook of the Philosophy of Science*. Elsevier, 2008. In press.