Advances in Cognitive Systems X (20XX) 1-6 Submitted X/20XX; published X/20XX

Simple Rules for Probabilistic Commonsense Reasoning

Adam Purtee APURTEE@CS.ROCHESTER.EDU
Lenhart Schubert SCHUBERT@CS.ROCHESTER.EDU
Department of Computer Science, University of Rochester, Rochester, NY, 14619 USA

Abstract

A long-sought goal of AI is the mechanization of human-like reasoning and argumentation based not only on firm knowledge supporting deduction but also generic conditional knowledge supporting tentative or probabilistic inferences. Arguably, most of the general knowledge people possess is of this type. We propose probabilistic "amplifying rules", "Bayes rules", and "categorization rules", combined with an algebraic conception of probability, as a basis for commonsense probabilistic reasoning, i.e., the generation of new, probabilistically qualified propositions from a commonsense knowledge base. Our emphasis in this paper is on amplifying rules (*a-rules*), and we report positive initial results when applying such rules to a classic Markov logic test set.

1. Motivating Rule-Based Uncertain Inference

Logical reasoning in propositional and quantified logics is based on rules such as *modus ponens* or natural deduction rules, which are easily understood intuitively, independently applicable, tolerant of knowledge elaboration, and semantically justifiable. But commonsense reasoning involves uncertainty, and so reasoning research has long been tantalized by the goal of generalizing traditional methods so as to allow for uncertainty. Here are some diverse examples of the sorts of unreliable, commonsense knowledge that people seem capable of using for inference:

1. Most dogs are friendly.
2. Dogs are usually well-treated by their owner.
3. New PhD graduates are usually under 30 years old.
4. Restaurants almost always serve alcoholic beverages.
5. Supermarkets rarely sell socks.
6. Smokers often have other smokers as friends.
7. Dogs are usually someone's pet.
8. If someone has a pet, it's quite likely to be a dog.
9. If someone has a pet, it's fairly likely to be a cat.
10. If someone hasn't eaten for several hours, s/he is probably hungry.
11. If a graduate student is among the authors of a research paper, his or her advisor is likely to be among the authors as well.
12. About 9 out of 10 research proposals are declined.

The point of this lengthy list of examples is to underscore just how heterogeneous our commonsense "rules of thumb" are, in terms of topical variety, propositional complexity, and reliability. Using very large knowledge bases containing such rules, taking account of the degree of certainty of the conclusions, lies well beyond current methods of uncertain inference in AI.

We should make a distinction here between general facts and rules. The first seven and last example are statements of fact, termed *generic sentences* in linguistic semantics. For our purposes, they can be viewed as rough statistical claims. The remaining examples (stated as conditionals) are best viewed as rules that directly suggest the degree of certainty of the consequent for any instance of the antecedent, when nothing else is known about that instance. However, the two kinds of formulations are closely related. For example, we can convert assertion (1) to a rule stating that *given any dog, it is quite likely to be friendly*. Conversely, we can regard rule (8) to be justified if in fact, say, 60% of pets are dogs. We will limit ourselves to rules that reflect statistical facts in this way. But note that as soon as we consider an *instance* of a rule, it no longer corresponds to a statistical fact, but rather is a rule for assigning a degree of belief to the consequent, when (the instance of) the antecedent is the only relevant knowledge.

An intelligent commonsense reasoning system should be able to accommodate this kind of knowledge (whether supplied as general facts or as rules), be able to communicate it, and be able to draw reasonable conclusions from it. As an example of the simplest kind of uncertain inference that a general AI systems should handle, suppose that the proportion of dogs that are friendly is known to be about .8 (cf. example 1), and all we know about Rover is that Rover is a dog. Then we can conclude that Rover is friendly, with certainty .8. The general form of this example is that we are given (only) that a certain x is of type A, and that a proportion p of As are Bs, and we conclude (nonmonotonically) that x is of type B with certainty p. (See examples 1-7, 12.) Such reasoning (sometimes called *direct inference*) seems utterly natural for people, and thus should be trivial for a commonsense reasoning system. Arguably human beings possess (and can approximately verbalize) many millions of knowledge items of this kind, where A and B may be logically complex, and p may vary by subtle degrees from *certainly* to *certainly not*. Qualitatively expressed certainties can be put into correspondence with numeric values only in rough ways, but it does matter that the degrees of certainty vary, and sometimes approximate numbers are known, as in the last example.

Of course, in general we have, or progressively acquire, diverse knowledge about an individual such as Rover, not just a single fact. As we take more and more of this knowledge into account, our confidence concerning a particular conclusion (such as Rover's friendliness) may shift up or down. Thus it is crucial to have sensible methods for *combining* inferences from miscellaneous facts bearing on the same conclusion. Also, our inferences do not in general consist of single steps. Rather, we draw conclusions of variable certainty from given facts (depending on the reliability of the rules employed), and proceed further from those conclusions, using any applicable rules. Thus it is also crucial to have sensible methods of *chaining* from uncertain propositions to further conclusions, appropriately assigning degrees of certainty to those further conclusions.

Our proposed *a-rules, b-rules*, and *c-rules* all allow for direct inference, and we specify well-motivated methods of combining and chaining inferences. Under certain conditions, certainties of inferences based on our rules can be obtained using just the numerical probabilities of the rule consequents, given the antecedents. This is not possible in general because of dependencies among premises and inference chains, but thanks to an important innovation in our probability calculations– the use of *algebraic probabilities*–we are able to take account of such dependencies in principle. We also propose a *sparse truth assumption* (STA) as a general assumption about "natural" predicates that is often tenable and can greatly simplify inference of numerical certainties for the conclusions reached.

Our main focus in this paper will be on a-rules, which are natural rules of positive evidence, always increasing the degree of certainty of their conclusion, whatever other evidence there might be. We suggest that a-rules often produce reasonable results even when the criteria for applying them are not strictly met, and we report experimental results on a well-known "students and their advisors" data set. The results produced by three a-rules are more accurate than the baseline results for the original Markov Logic Network approach for which the data set was intended.

2. Related Work

Various approaches to uncertain inference have been developed over the past decades, but they fall short of generalizing classical reasoning in a way that naturally extends such reasoning, retaining its attractive properties while allowing for degrees of certainty.

Production systems (also called rule-based systems or expert systems) such as Shortliffe and Buchanan's pioneering MYCIN system for diagnosing infections and recommending therapy (Shortliffe and Buchanan, 1975) use intuitive if-then rules with bounded numerical "certainty factors" (e.g., in the interval [-1.0,1.0]); these are combined in ways intended to boost or lower certainties in an intuitively natural direction, without going out of bounds. However, the rules operate on attribute-value lists, rather than logical formulas, and the certainty factors are not grounded in statistical knowledge or combined in ways that have any theoretical basis.

Bayesian networks (BNs) have the important advantage that known conditional frequencies relating rule antecedents to consequents can be used directly to set the network parameters; as well, they exploit independence assumptions that often seem justified, for example when the node-to-node connections can be interpreted as causal influences. But BN inference is restricted to propositional variables, and involves complex marginalization processes (often approximated by sampling methods) quite unlike classical reasoning. While various Prolog-like quantified extensions exist Milch et al. (2007); Raedt et al. (2007); Getoor and Grant (2006), uncertain inference generally devolves into performing standard BN inference on *ad hoc* BNs built from ground instances of quantified relationships.

Undirected graphical models such as Markov networks share the advantage of BNs that their parameters are anchored in empirical frequencies. An advantage they have over BNs is that they do not presuppose any causal knowledge – "neighboring" variables are simply regarded as statistically related. However, the flip side is that the network parameters cannot be set directly using conditional frequency knowledge, but rather must be mathematically inferred from large data sets. This is a severe limitation, in view of the abundance of commonsense knowledge that we would like to impart to machines. Also, as in the case of BNs, quantified versions of undirected graphical models such as Markov Logic Networks (MLNs) rely on ad hoc grounding to perform uncertain inference, and there is no provision for the simple kind of direct inference about the likelihood of x being a B, given that it is an A, mentioned above.

Finally, formal extensions of first-order logic allowing for statistical knowledge and probabilistic qualification of sentences tend to be strong on semantics but weak on inference mechanisms. The probability of a conclusion is typically measured in terms of the proportion of possible worlds where it holds, but calculating this proportion (perhaps in the limit as the domain of individuals is expanded) can be extremely challenging, and this calculation must be performed each time the kb is expanded. Bacchus *et al.* proved in (1996) and (1993) that first-order probabilistic reasoning based on model counting has desirable theoretical properties, such as respecting specificity and entailment relations, and supporting direct inference, but exact inference for knowledge bases with more than monadic predications is intractable. More recently, promising work has been done on lifted inference which aims to provide tractable algorithms for inference with model counting semantics (Gribkoff et al., 2014; Kazemi et al., 2016), but these methods still fall short of classical style reasoning with first-order information.

Our work is aimed at providing a probabilistic extension of quantified logic, similar in style to default logics (and to some extent production systems and BNs) but using rules that

- are framed in terms of logical antecedent and consequent formulas, allowing for matchable variables and for uncertainty of the consequent;
- are grounded in empirical frequencies (perhaps just estimates from linguistically stated generalizations); and
- are independently applicable (under certain provisos) in drawing uncertain conclusions from a given set of premises.

Further, we seek rules that jointly yield sensible results in cases where we know what the results should be, such as in Bayesian networks (BNs). In the rest of the paper, we discuss *a-rules* in some detail, motivating them and considering their range of applicability. Our methods allow rule-based inferences in Bayesian networks that agree with those grounded in standard BN theory, and they provide intuitively reasonable, quantitative counterparts to nonmonotonic reasoning methods. As mentioned in the introduction, our main experimental test shows that a straightforward application of three a-rules can more than match the inferential accuracy of Markov Logic Networks, as originally tested by Richardson and Domingos (2006) in a “graduate students and their advisors” domain.

3. Rules

We write rules as implications with a *certainty variable* modifying the consequent (or in c-rules, multiple certainty variables, each modifying one of a set of alternative consequents). Certainty variables are treated as boolean random variables (Bernoulli variables) whose probability is the probability of truth of the consequent, when the antecedent is true. Therefore, a certainty variable can be thought of as added conjunctively to the rule antecedent, so that the consequent is true whenever both the antecedent and the certainty variable are true. This probability should at the same time reflect the proportion of true instances of the antecedent where the consequent is true.

We should note here that a certainty variable is a unique Bernoulli variable only in a rule that contains no matchable ("universal") variables as predicate arguments. When there are such matchable variables, these are also considered arguments of the certainty variable, i.e., the latter is in general a *function* of the matchable rule arguments. The value of this certainty function is regarded as a distinct Bernoulli variable for each combination of values of its arguments. For example, given a rule stating that an arbitrary dog x is friendly with probability $(p\ x)$ (see below), $(p\ \textit{Rover})$ and $(p\ \textit{Fido})$ are distinct random variables, assuming that the names refer to distinct individuals.

In general, rules used for uncertain inference should be *stable* in the sense that their certainty variables should have fixed probabilities, rather than ones that shift in the course of a reasoning process, as a result of shifts in the certainties of various propositions affected by the reasoning process. Further, the overall effect of applying a set of rules should be independent of their order of application. (However, we found that we needed to relax that constraint somewhat in considering the interaction among rules leading to contrary conclusions.)

The intuitions behind our three types of rules are quite different; accordingly, they differ in the kinds of knowledge they can be used to encode, and in the way they update consequent probabilities (given the truth or at least the probability of their antecedents), i.e., their combinatory and chaining behavior. Here, we discuss a-rules in some detail, while only briefly characterizing b-rules and c-rules.

3.1 a-Rules

Amplification rules capture a simple, intuitive notion of some event or predication justifying an increased belief in some other event or predication. For example, consider the following simple a-rule:

$(\forall x[[x\ dog] \Rightarrow (:a\ (p\ x)[x\ friendly]])$,

i.e., for any dog x, conclude with certainty $(p\ x)$ that x is friendly (in the absence of other information). Here $\forall$ is being loosely used not as a universal quantifier, but as an indication that x is a matchable rule variable.[1] Note also that we are using square brackets and predicate infixing for sentential formulas, and round brackets and functor prefixing for functional expressions. As noted above, $(p\ x)$ is a distinct Bernoulli variable for each value of x. Moreover, these Bernoulli variables are considered independent of one another and of the certainty variables associated with other a-rules. (As such they are *choice variables* in the sense of Poole (2008).)

We noted above that the Bernoulli random variables $(p\ Rover)$ and $(p\ Fido)$ are treated as independent of one another. However, we normally take the instantiated certainty variables of any given rule to have the same probability of truth. For example, using vertical bars to indicate numerical probabilities of random variables, we may write

$|(p\ x)| = .8$ for all x,

so that the rule in effect affirms that any arbitrarily selected dog is 80% likely to be friendly. Note that we would assume this rule to be grounded in an empirical claim, namely, that 80% of dogs are friendly. As explained earlier, once the rule has been instantiated, it no longer reflects any statistical fact, but is just a rule for assigning or updating the certainty of the consequent, given the antecedent.

1. We misuse '`all`' analogously for rules in the EPILOG inference engine, for convenience in the implementation.

This kind of rule instantiation immediately provides a means of direct inference of probabilistically qualified conclusions – something unavailable in, for example, NMR methods and MLN methods.

Direct inference applies only if no other knowledge has (yet) been applied to the conclusion. We started by saying that a-rules serve to increase belief in a conclusion; so suppose that the certainty of a conclusion has already reached some level p in a reasoning process, leaving remaining uncertainty $(1-p)$. Then an a-rule with certainty variable q will add a fraction q of the remaining uncertainty $(1-p)$ to the certainty p, with result $p + q.(1-p)$. (To take account of a possible dependency of p on q, we replace multiplication of numbers by idempotent multiplication of algebraic probabilities below.) Note that a-rules can be arbitrarily weak or strong (with $0 < |q| < 1$), but a sufficient number of weak rules may push the certainty of a conclusion arbitrarily close to 1.

Also note that direct inference corresponds to applying an a-rule in a case where the prior probability of the conclusion is vanishingly small (i.e., with $p = 0$, the update to $p + q.(1-p)$ is precisely q). We believe that ascribing near-0 probability to ground predications, in the absence of any relevant knowledge about the arguments of the predicate, is often very reasonable. We call this the *sparse truth assumption* (STA), in recognition of its similarity to the *closed world assumption* (CWA) often used in nonmonotonic reasoning. The plausibility of the STA can be appreciated, for example, by considering what proportion of entities in any general world ontology are *dogs*, or are *friendly*, or *serve alcoholic beverages* (see our previous examples)—surely, essentially 0, in view of real-world entities such as stars or grains of sand, let alone abstract ones like numbers.

The combinatory behavior of a-rules, apart from being intuitively natural, can be further motivated in the following way. If certain distinct facts provide *independent support* for a conclusion, then the certainty of the conclusion should behave like the probability that independent Bernoulli trials will yield "success", i.e., a positive outcome on at least one of the trials. This intuition leads to the *noisy-OR* rule of combination,

Pr(success) = $1 - (1-p)(1-q)(1-r)...$,

where $p, q, r, ...$ are the (magnitudes of) the choice variables of the rules whose antecedents independently lend support to the shared conclusion. (For 2 variables this is $p + q - p.q$.) A classical example is "Mr. Holmes' burglar alarm", which may be triggered by an intruder or an earthquake; certainly, if either (or both) of these events occurred, they independently suggest that the alarm may have been triggered (Pearl 1988:49-50). A key observation for our purposes is that the noisy-OR result can be obtained with separate rule applications, in any order; viz., we simply add an increment to the current certainty of the conclusion, where that increment is the choice-variable probability p times the amount by which the current probability falls short of 1.0. So if the initial certainty is 0 (the conclusion is extremely improbable, in the absence of evidence for it), then updating with rule probability (choice-variable probability) p yields certainty p for the conclusion. If we now update independently with a rule probability q, we obtain certainty $p + q(1-p) = p + q - pq$, i.e., the noisy-OR value, for the conclusion.

As may already be clear from our examples and discussion, the general form of an a-rule is

$(\forall x_1 ... x_k\ [\phi \Rightarrow (\text{:}a\ (p\ x_1 ... x_k)\ \psi)])$,

presumed to involve (match) variables $x_1, ..., x_k$. The distinct values of the match variables correspond to distinct, independent Bernoulli random variables $(p\ x_1 ... x_k)$. In implementing a-rules, we represent a hypothesis H obtained by uncertain inference as a certainty-modified statement such as $(:a\ p\ H)\}$. Treating this as an initial knowledge base Δ, we can represent the operation of adding another such statement about H based on another a-rule as $\Delta' = \Delta \cup \{(:a\ q\ H)\}$, where $Pr(H|\Delta') = p + q * (1 - p)$, and the '$*$' operator represents idempotent algebraic product, discussed below.

Noisy-OR belief combination has been widely used in applications where all the influences on the certainties of the propositions (domain variables) of interest are believed to be causal influences. For example, large 2-level BNs have been constructed in which diseases are roots (at level 1) and the "findings" at level 2 can be caused by one or more diseases (or by an independent unknown cause). It is hardly surprising that causal models have played such a prominent role in uncertain inference, since in some sense the quest for understanding the phenomena in the world around us, including illness, is a quest for discovering underlying causal mechanisms. As well, much evidence has been accumulated that human perception and cognition involve causal Bayesian inference (Gibson et al., 2013; Jacobs and Kruschke, 2011).

Now, some of the sample rules we have listed are arguably causal, even if the causal mechanisms are obscure. For instance, we could say concerning claim (1) that it is the intrinsic nature of dogs that generally causes them to be friendly; or, concerning (2), that the intrinsic nature of dogs and dog owners generally causes the owners of the dogs to treat them well. But we could hardly claim, concerning (3), that the intrinsic nature of new Ph.D. graduates causes them to be under 30 years old. Various other rules in our list resist a causal interpretation.

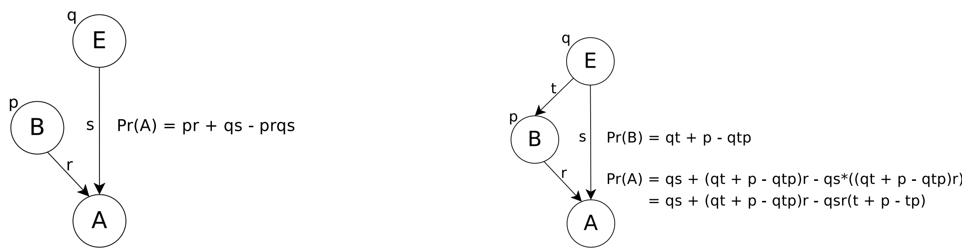
Nonetheless, we hypothesize that we can cast many generalizations as a-rules, even if they resist a causal interpretation. In a supplementary document (Schubert, 2017) we cast (3) as an a-rule and apply this in combination with another a-rule to the effect that a person with a young sibling is apt to be young as well. We find that the certainty that a new Ph.D. with a 24-year-old sibling is under 30 aligns well with the results of a detailed generative model for age differences between siblings, and ages for attaining a Ph.D. The main evidence we bring in this paper for the efficacy of a-rules is the set of results for the "graduate students and their advisors" domain.

3.2 Algebraic Probabilities

Before discussing b-rules and c-rules, and in preparation for fuller description of our inference methods, we need to provide a quick characterization of algebraic probabilities.

Algebraic probabilities were introduced in (Schubert, 2004) (under the name "quasi-probabilities") as a general means of manipulating and evaluating probabilities in Bayesian networks (BNs) built from noisy-AND/OR/NOT nodes – which can model all boolean-valued BNs. The combinatory algebra of these probabilities is closely analogous to ordinary algebra using +, -, and product, except that product, written '*', is *idempotent*, i.e., $\alpha * \alpha = \alpha$ for any algebraic probability expression α. (The idempotency of '*' derives from that of logical $\wedge$, while $(1 - \alpha)$ represents the probability of a negation). Expressions are formed by combining *elementary* probabilities, which are independent of one another. For a product $\alpha * \beta$ where α and β share no elementary probabilities, '*' reduces to ordinary product, i.e., $\alpha * \beta = \alpha\beta$.

A fundamental problem in uncertain inference is that inference chains leading to the same (or opposite) conclusions may rely on some of the same uncertain knowledge items along the way. In such a case the two inference chains do not provide independent evidence for the conclusion, and so



(a) Standard (b) Modified

Figure 1: The Burglary Alarm Earthquake example.

the probability of the conclusion cannot be obtained correctly by combining the numerical results of the inference chains as if the chains were independent. Rather, we need to take account of the "provenance" of the results being combined. Algebraic probabilities provide a general way around this difficulty, at least in principle.

The simplest demonstration of this point is provided by considering applying the same rule twice to the same premise, with the same conclusion. In such a case, the two rule applications are certainly not providing independent evidence, but rather the *same* evidence, twice. Yet by using algebraic probabilities and idempotent products we obtain the correct result. Suppose that when the rule is first applied, the probability of the conclusion is still 0. Then if the certainty variable of the rule is p, the result will be p. If the rule is applied a second time, the result will be $p+p*(1-p)$ = $p + p - p = p$, i.e., unchanged. It is easy to show that this still holds if the initial probability of the conclusion is some arbitrary algebraic value α.

A more subtle example is provided by a variant of "Mr. Holmes' burglar alarm". The standard version is shown in Figure 1(a), where a burglary B and an earthquake E are independent causes of an alarm A (and they are the *only* possible causes). In variant (b), it is assumed that burglars may take advantage of earthquakes to commit burglaries. Thus both influences on A have a common ancestor in their possible causation, E. As such, they are no longer independent. Yet, as is shown in the figure, when we combine the influences "as if" they were independent – but using idempotent product – we obtain the exact result for the probability of A (as the reader can verify).

Details of algebraic probability manipulation in BNs can be found in (Schubert, 2004), but we should mention the general conditioning rule: For any prior algebraic probability $Pr^*(\phi)$, when evidence ψ is brought to bear, the resultant probability is

$$Pr^*(\phi)\frac{*Pr^*(\psi)}{*Pr^*(\psi)} \ ,$$

where $Pr^*(\phi)$ denotes the algebraic probability of ϕ (similarly, $Pr^*(\psi)$) and the two occurrences of $*Pr^*(\psi)$ indicate idempotent multiplication of the numerator and denominator of $Pr^*(\phi)$ by $Pr^*(\psi)$. For example, suppose that A is given as evidence in Figure 1(b). Then the probability of E is updated to

$$\frac{q*[(p+qt-pqt)r+qs-qsr(p+t-pt)]}{(p+qt-pqt)r+qs-qsr(p+t-pt)} = \frac{q[(p+t-pt)r+s-sr(p+t-pt)]}{(p+qt-pqt)r+qs-qsr(p+t-pt)} = \frac{q[s+r(1-s)(p+t-pt)]}{(p+qt-pqt)r+qs-qsr(p+t-pt)}.$$

A convenient generalization of the algebra, for example in handling c-rules or mutually exclusive rule antecedents, is to allow "spectra" $p_1, p_2, ..., p_k$ of elementary probabilities, where $p_i * p_j$ = 0 for distinct i, j. Note that this generalizes the fact that $p*(1-p)$ = $p-p$ = 0. Members of different elementary spectra are considered independent of one another.

3.3 b-Rules

The best example of *b-rules* (Bayes rules) are rules that duplicate the effect of naïve Bayesian inference from evidence items to a hypothesis, where the evidence items are conditionally independent of one another, given the hypothesis. Naïve Bayesian inference is often used as an approximate method of uncertain inference in reasoning from effects to their causes, despite its weaknesses. For example, suppose that an effect has two possible alternative causes, and we learn that the effect is true, and update the probabilities of the causes accordingly. If we then learn that one of the causes is true, the probability of the other cause should drop – the effect has been "explained away", so that a second explanation is unnecessary or at least less likely. Naïve Bayesian inference fails to account for this phenomenon.

Nonetheless, naïve Bayesian inference is sometimes correct, so we would like to cast Bayesian rules in a form such as

$(\forall x[[x \textit{ has-runny-nose}] \Rightarrow (:b\ (p\ x)[x \textit{ has-cold}]])$.

Now, it is well-known that naïve Bayesian inference can be performed one evidence item at a time by using likelihood-ratio updating of the odds in favor of the conclusion (hypothesis) in question. For example, we could update the odds in favor of x having a cold by multiplying the prior odds of that eventuality by the likelihood ratio

$$\frac{Pr([x\ has-runny-nose]|[x\ has-cold])}{Pr([x\ has-runny-nose]|\neg[x\ has-cold])}.$$

Thus we could supply such likelihood ratios as b-rule parameters. However, for uniformity we wish to use probability parameters in all rules; so $(p\ x)$ should lie between 0 and 1. Therefore in general, instead of supplying $Pr(E|H)/Pr(E|\neg H)$ as b-rule parameter for a rule $(E \Rightarrow (:b\ p\ H))$, we instead use

$p = Pr(E|H)/[Pr(E|H) + Pr(E|\neg H)]$.

It is easily verified that with this choice of parameter, $p/(1-p)$ is the likelihood ratio needed for updating the odds in favor of H. Note that applying a b-rule is thus quite different from applying an a-rule: Instead of using the certainty variable to *increment* the probability of the consequent, we are using it to *scale* the odds, and thus in effect the probability, of the consequent. This scaling may either raise or lower the probability of the consequent, depending on whether the likelihood ratio is greater or less than 1.

We have already pointed out that naïve Bayesian inference has problems with "explaining away". The way this shows up in formulating b-rules is that the p-parameter of a rule $(E \Rightarrow (:b\ p\ H))$, when there is also an alternative rule $(E \Rightarrow (:b\ p'\ H'))$, will involve the prior probability of H'. But since this probability is itself subject to change as a result of inference, the b-rule will no longer be stable – it parameter may well change as other inferences are made.

Because of this defect, b-rules seem to have limited applicability in general reasoning. Interestingly, by reliance on algebraic probabilities we could theoretically do *all* reasoning using b-rules.

(This is a reflection of the fact that Bayes' rule is universally valid.) But such an approach would be equivalent to the general conditioning method for algebraic probabilities mentioned above it – and like that method, would require computing algebraic marginals for all propositions of interest.

3.4 c-Rules

Categorial rules are intended to encode conditional knowledge where truth of the antecedent redistributes likelihoods among a competing spectrum of alternative (and exhaustive) possibilities. For example, learning that a person that we know little about is a grandparent may shift our beliefs about the person's likely age category, and knowing that the person has a 30-year-old sibling may shift them in another direction. Other clear examples of this arise when reasoning with taxonomic information, e.g., in guessing whether the animal that chewed open some plastic garbage bags overnight is a dog, raccoon, skunk, bear, or something else. We write c-rules in the form (neglecting matchable predicate arguments for simplicity here),

$E \Rightarrow (:c\ q_1\ H_1\ q_2\ H_2\ ...q_k\ H_k)$.

Like b-rules, c-rules can be understood in terms of likelihood ratio updating. If disjoint hypotheses $H_1, H_2, ..., H_k$ each suggest the truth of E with some likelihood $Pr(E|H_i)$, $i = 1, ..., k$, then the prior probabilities of the H_i can be updated by multiplying them by those likelihoods, in effect forming a Hadamard product, which is then normalized by dividing by the sum of the individual product terms. Since we normalize after applying a c-rule, we could simply use $q_i = Pr(E|H_i)$ as the certainty variables, but we wish to view the q_i as a distribution over the H_i when we know only E. Therefore we normalize the q_i as $q_i = Pr(E|H_i)/\sum_j Pr(E|H_j)$. Algebraically, we treat the q_i as a spectrum of mutually exclusive Bernoulli random variables, independent of other such spectra and of certainty variables in a-rules. Thus, if the prior probabilities of $H_1, H_2, ..., H_k$ are $p_1, p_2, ..., p_k$ respectively, then with q_i as just defined, the updated probabilities, given E, are

$p'_1, p'_2, ..., p'_k$,

where $p'_i = p_i q_i / Z$, for $i = 1, 2, ..., k$, $Z = \sum_{i=1}^k p_i q_i$.

Notably, this update rule coincides with Dempster-Shafer updating when the D-S "frame of discernment" consists of disjoint sets and the prior distribution over those sets is uniform.

As in the case of a-rules, we find it generally appropriate to apply the STA, i.e., the H_i all have vanishingly small probabilities prior to application of known facts and rules (and a vanishingly small sum of these priors); but once E is affirmed, the posterior probabilities of the H_i add up to 1. In other words, the H_i categories are exhaustive, given E.

It is also worth noting the connection to b-rules. Any b-rule $E \Rightarrow (:b\ q\ H)$, where q = $Pr(E|H)/[Pr(E|H) + Pr(E|\neg H)]$ (see previous subsection) could be reformulated as a c-rule $E \Rightarrow (:c\ q\ H\ q'\ \neg H)$, where q = $Pr(E|\neg H)/[Pr(E|H) + Pr(E|\neg H)]$. However, in this case the specified alternatives do not have vanishingly small probabilities before knowledge is applied – rather, $Pr(H)$ is near 0, and $Pr(\neg H)$ is near 1.

Because c-rules simultaneously take account of interactions among alternative conclusions, given the antecedent, they are much more broadly applicable than b-rules. In fact, they can be used in principle to reason "anti-causally" from effects to possible causes in BNs. For example, suppose that in a BN for, say, disease diagnosis, a certain symptom E has three possible causes A, B, C (and only these). Assume that they influence E in noisy-OR fashion. Then we can formulate a c-rule whose disjoint conclusions correspond to the possible combinations of truth values of A, B, C, so that there will be 7 categories (with at least one of A, B, C true):

$E \Rightarrow (:c\ q_a\ \{A\}\ q_b\ \{B\}\ q_c\ \{C\}\ q_{ab}\ \{AB\}\ ...\ q_{abc}\ \{ABC\})$.

where $\{A\}, ..., \{AB\}, ..., \{ABC\}$ respectively denote the hypotheses that *only* A is true, ..., the hypothesis that *only* A and B are true, ..., and finally that all three of A, B, C are true. Writing the single-cause probabilities as $Pr(E|\{A\}) = u, Pr(E|\{B\}) = v, Pr(E|\{C\}) = w$, we can rewrite the c-rule in unnormalized form as

$E \Rightarrow (:c\ u\ \{A\}\ v\ \{B\}\ w\ \{C\}\ u + v - uw\ \{AB\}\ ...\ 1 - (1-u)(1-v)(1-w)\ \{ABC\})$.

Having applied such rules, we can find individual posterior probabilities of causes A, B, C by marginalizing, e.g., adding the probabilities of the combinations $\{A\}, \{AB\}, \{AC\}, \{ABC\}$ in which A is true in order to obtain the posterior probability of A.

For example, such rules can completely model inference of root causes in any 2-level noisy-OR BN, given the truth of some of the symptoms. (Symptoms known to be false require separate rules.) Of course, the sizes of the c-rules grow exponentially with the number of causes a symptom can have. On the other hand, if all of the causes C have very low marginal probabilities compared to their single-cause probabilities $Pr(E|C)$ (i.e., the STA holds), then the logical combinations of causes where more than 1 or 2 are true may well have negligibly small probabilities, i.e., the symptom E is very unlikely to have more than 1 or 2 causes. In the above example, if we can neglect all but the first-order terms, the (unnormalized) rule just becomes

$E \Rightarrow (:c\ u\ A\ v\ B\ w\ C)$,

i.e., it is as if A, B, C were mutually exclusive alternatives. We have not yet experimented with exact or approximate c-rules, but conjecture that together with a-rules they will provide a very powerful basis for general rule-based reasoning.

4. Inference

We have implemented inference with a-rules through extensions to a first-order theorem prover based on the language of Episodic Logic (Schubert and Hwang, 2000). Our inference algorithms are built upon extensions of logical deduction – though we emphasize that they are not strictly deductive rules, but instead use deductive reasoning machinery to combine evidence.

4.1 Rule Combination

Two fundamental challenges of reasoning with a-rules pertain to combining the effects of convergent rules (i.e., with a shared conclusion) and reasoning with the resultant uncertain conclusions. In general, we combine convergent a-rules with an algebraic implementation of noisy-OR. This means that the effect of an a-rule will always be to amplify the likelihood of its consequent, regardless of the magnitude of the corresponding algebraic probability. It is straightforward to do this mechanically for an arbitrary first-order formula ϕ:

From $(:a\ p\ \phi), (:a\ q\ \phi)$ infer $(:a\ (p\ +\ q\ -\ p * q)\ \phi)$

Since the product $p * q$ is algebraic, identical terms are multiplied idempotently, and mutually exclusive terms yield zero. The algebraic approach robustly allows for arbitrary a-rule combinations (even with themselves, as already noted).

The complementary problem of reasoning with uncertain conclusions is similarly tackled with a combination of deduction-style rules and algebraic probabilities. Suppose that the antecedent ϕ of a rule $\phi \Rightarrow (:a\ q\ \psi)$ becomes established with certainty p (algebraically expressed). Then, since the choice variable q is in effect conjoined with ϕ, the probability of the conclusion is $p * q$. In other words, rule chaining is a matter of (idempotently) multiplying the antecedent probability by the choice-variable probability. Technically, we obtain this result with a chaining rule

From $\phi \Rightarrow \psi, (:a\ p\ \phi)$ infer $(:a\ p\ \psi)$,

where we allow ψ to be replaced by $(:a\ q\ \psi)$.

As an example, if most dogs are friendly animals, and most friendly animals make good pets, then a given dog is rather likely to make a good pet. Further, if the dog is known to be a labrador, and we have no more specific knowledge about the friendliness of labradors than that of dogs, we would still make the same inference.

4.2 Negative, Disjunctive, and Existential Rule Consequents

a-Rules with negative conclusions need to be distinguished from those with positive conclusions, since under the STA, the knowledge-free probability of a negative conclusion is near-1 rather than near-0. So we track a-rule applications leading to a positive predication separately from ones leading to its negation. We combine their effects by treating the negative conclusion as a "spoiler" for the positive conclusion; for example, if the positive conclusion receives probability p and the negative conclusion receives probability q, then the resulting probability is $p * (1 - q)$. As an intuitive example, sunshine and warm air might both suggest pleasant weather, but their combined effect will be spoiled if strong winds are present also, and tend to imply unpleasant weather.[2]

Reasonable handling of disjunctive rule-consequents depends on the presumed relationship among disjuncts. If they are disjoint, the appropriate formalization is in terms of c-rules. If they can be considered conditionally independent, given the antecedent, then it makes sense to split the rule, also splitting the choice variable p of the rule into independent elementary probabilities $p_1, ..., p_k$, by default of equal magnitude, say x, satisfying $|p| = 1-(1-x)^k$ (so that the numerical probability of the disjunction with come out to $|p|$). If possible, however, one would use empirical data to assign approximate values to the $|p_i|$.

We have not yet adequately explored interactions between existentials and a-rules, either analytically or empirically. One possibility is Skolemization, with due attention to the fact that a Skolem constant may well share its denotation with a proper name ("*An AI professor* advises Bill, namely Mary"). There are also issues of the relative scopes of probabilistic qualification and existentials ("Probably an AI professor advises Bill", vs. "There is an AI professor who probably advises Bill".)

4.3 Obtaining Final Probabilities

Our algorithm for obtaining final probabilities relies on three stages: construction of the inference graph, deriving algebraic probability expressions from the graph, and then numerically evaluating the probabilities.

2. Here is a sanity check on our rule of combination: Suppose we regard the probability $(1 - p)$ of $\neg\phi$ suggested by the positive evidence for ϕ as evidence against ϕ. Substituting this for q in $p * (1 - q)$ gives back p – as it should!

We construct the inference graph by forward reasoning. Since our empirical comparisons herein are concerned with finite Markov Logic kbs, we were able to use exhaustive forward inference.[3] This method allows us to distinguish evidence-based uncertainty (viz., when an a-embedding can be retrieved) and simple ignorance (retrieval failure). Also in contrast with Markov Logic, we can compute probabilities incrementally through local updates to the graph as new information is discovered.

Our algorithm does not descend recursively into quantified contexts, but because our inference algorithms are first-order, derivable relationships between quantified wffs will in principle be correctly accounted for. We leave fuller investigation of this topic to future work.

After construction of the inference graph (or a subset), we obtain algebraic probability expressions for arbitrary formulas based on a recursive querying procedure. If a formula (or its negation) is known to be true without probabilistic qualification, then the probability is simply 1 (or 0). If a formula is a literal, then we separately obtain the embedding a-statements for it and its negation, and combine them as outlined above. For conjunctions about which we have no direct knowledge, we form the algebraic product of the algebraic probability of each conjunct. Finally, for disjunctions, we form the algebraic noisy-OR of the algebraic probabilities of the disjuncts.

Ultimately, we numerically evaluate the determined algebraic probability of the formula in question. When all of the associated algebraic probabilities are independent, this reduces to simple numerical substitution and arithmetic evaluation. When algebraic probabilities are repeated (or found together with their complements), then we must first eliminate the redundant and conflicting symbols, which can be computationally complex. In general, this process is NP-complete. Numerical evaluation of algebraic probabilities can be computationally complex, depending on the relationships between variable bindings. In the worst case, it is NP-hard, as is proved by by reduction from 3SAT (Schubert, 2004); however, this can be improved for common cases through resolution-like simplifications and common factor extraction.

5. Experiments

In this section, we present empirical evidence in support of the validity of our method by predicting advisor-advisee relationships over a moderately large database of facts about a computer science department. This dataset was introduced to the community by Richardson and Domingos (2006), who provide a strong baseline using Markov Logic; hereafter, we refer to it as the UW-CSE dataset.

A Markov logic network (MLN) is a set of weighted first-order statements. Inference in MLNs consists of constructing a large Markov random field with nodes for each ground predication. The probability of a model (assignment of truth to all ground formulas) is given by a normalized exponential weighting function. Inference of the probability of formula is performed by marginalizing over all worlds wherein the formula is true. Markov logic is a widely-studied field with many applications. Many algorithms exist for inference and learning, and most rely on some form of statistical sampling. Exact inference in Markov logic is #P complete in the size of the domain.

3. In future we may attempt using simplification of algebraic probabilities to detect convergence (as fixed points), and use "interestingness" and low-probability criteria to limit forward inference, as was done in a previous version of the EPILOG inference engine – which, however, computed probabilities in a very ad hoc way (Schubert and Hwang, 2000).

In this section, we compare accuracy and computational demands of our model that obtained with the original MLN. In the following years, structure learning and improved inference algorithms boosted the performance of Markov logic networks on this dataset; however, we assess our method with respect to the original work.

5.1 Setup

The UW-CSE dataset consists of two main components:

- A database of ground facts describing academic relationships among professors, students, courses, and publications within a computer science department. The set of terms ranges over persons, publications, and courses. The set of predicates include type information (person vs publication), publication authorship, course instructors, teaching assistant roles, and advisor-advisee relationships. The complete dataset includes 1320 constants and 34 predicates and relations. The constants are partitioned into non-overlapping subsets by sub-area of computer science (ai, systems, theory, languages, and graphics.)

- The UW-CSE knowledge base is provided in two formats. At its core is a first-order knowledge base constructed by surveying members of the department for simple natural language statements describing the department which was then semi-manually edited for first-order syntax. The dataset also includes an MLN representation of the same first-order knowledge base. The baseline Markov logic network includes 82 rules.

To obtain baseline results, we use the MLN supplied by the authors of the dataset as well as the publicly available Alchemy implementation.

To apply our method, we converted all first-order implications found in the supplied knowledge base into amplification rules. Because MLNs make a unique names assumption, we add an additional $O(n^2)$ predications to each subset enforcing that assumption (using negated SameName, SameCourse, SamePublication, etc., predicates). This expansion of the database is largely responsible for computational complexity in our method as applied to the dataset. As a way of minimizing the number of rules used in evaluating our method, we omitted all first-order statements that do not immediately involve the target relation. This leaves us with three rules regarding co-teaching and co-publishing.

- $(\forall s, c, p, q$
  $(((s\ \mathit{Phase}\ \mathit{postQuals})\ \wedge\ (c\ \mathit{TaughtBy}\ p\ q)\ (c\ TA\ s\ q)\ \neg(c\ CourseLevelLevel 100)$
  $\Rightarrow (\text{:}a\ (P1\ s\ c\ p\ q)\ (s\ \mathit{AdvisedBy}\ p)))))$
- $(\forall s,\ c,\ p,\ q$
  $(((s\ \mathit{Phase}\ \mathit{postGenerals})\ \wedge\ (c\ \mathit{TaughtBy}\ p\ q)\ (c\ TA\ s\ q)\ \neg(c\ CourseLevelLevel 100))$
  $\Rightarrow (\text{:}a\ (P2\ s\ c\ p\ q)\ (s\ \mathit{AdvisedBy}\ p))))$
- $(\forall t,\ p,\ s$
  $(((t\ \mathit{Publication}\ p)\ \wedge\ (t\ \mathit{Publication}\ s)\ \neg(p\ SamePerson\ s)\ (p\ Professor)\ (s\ \mathit{Student}))$
  $\Rightarrow (\text{:}a\ (P3\ t\ p\ s)\ (s\ \mathit{AdvisedBy}\ p))))$

We obtain numerical parameters for our a-rules by simply counting conditional frequencies in the dataset. We perform inference using our model as outlined previously in this paper. Following

Setup	NRules	Learn (m)	Infer (m)	AUC
Full MLN	82	9.1	6.5	0.320
Small MLN	3	0.5	0.1	0.227
a-rules	3	45	122	0.370

Table 1: Computation time and mean area under the precision-recall curve when predicting the AdvisedBy relation between pairs of people. Generative learning was used for the MLN setups. All experiements were performed using the same system, and with minimal external load. The system has a core i7-2600K Intel processor and 16GB available RAM. The MLN uses 14GB ram while our method uses less than 2GB.

Richardson and Domingos, we report area under the precision-recall curve (AUC) results obtained by cross-validation across the subareas of the dataset in Table 1.

5.2 Results

Using our method, we obtain more accurate recovery of advisor-advisee relationships than the baseline MLN method; however, this comes at a significant computational expense. As noted this expense stems from the explicit world closure predicates (e.g., SamePerson, SameCourse, SamePublication). On average, this causes our stored database to grow from approximately 250 type predications to 8200 per subset. In principle we could obtain a dramatic speedup by using a sparse representation for some predicates. We also emphasize that unlike MLNs, our method is well-suited for open domains. Similarly, our parameter learning method is based on a simple recursive descent computation, which can in principle be performed much more efficiently through careful use of hashtables and conjunct ordering.

6. Conclusion

We have developed a novel method of integrating probability as belief with first-order logic. Our method is aimed at the kind of general knowledge readily expressed through language, and employs intuitively plausible, quantitative methods of evidence combination. Our deduction-style rules for reasoning with algebraically defined probabilities provide a step-wise proof theory while assuring that evidence won't be double-counted. We provide empirical evidence in support of our method through a comparison with Markov Logic in a relation prediction domain, where we employ simple conditional frequencies as rule parameters and obtain better accuracy scores. Immediate future work includes empirical investigation of categorial rule performance, as well as direct comparison to Bayesian methods as outlined in the related work.

Acknowledgements: This work was supported by Communicating with Computers, DARPA subcontract W911NF-15-1-0542. We thank the anonymous referees for their helpful comments.

References

Bacchus, Fahiem, Adam J Grove, Joseph Y Halpern, and Daphne Koller. 1993. Statistical foundations for default reasoning. *IJCAI*.

Bacchus, Fahiem, Adam J Grove, Joseph Y Halpern, and Daphne Koller. 1996. From statistical knowledge bases to degrees of belief. *Artificial intelligence*, 87(1):75–143.

Getoor, Lise and John Grant. 2006. Prl: A probabilistic relational language. *Machine Learning*, 62(1-2):7–31.

Gibson, Edward, Leon Bergen, and Steven T. Piantadosi. 2013. Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, 110(20):8051–8056.

Gribkoff, Eric, Guy Van den Broeck, and Dan Suciu. 2014. Understanding the complexity of lifted inference and asymmetric weighted model counting. *CoRR*, abs/1405.3250.

Jacobs, Robert A and John K Kruschke. 2011. Bayesian learning theory applied to human cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(1):8–21.

Kazemi, Seyed Mehran, Angelika Kimmig, Guy Van den Broeck, and David Poole. 2016. New liftable classes for first-order probabilistic inference. *CoRR*, abs/1610.08445.

Milch, Brian, Bhaskara Marthi, Stuart Russell, David Sontag, Daniel L Ong, and Andrey Kolobov. 2007. Blog: Probabilistic models with unknown objects. In *Statistical relational learning*.

Poole, David. 2008. The independent choice logic and beyond. In *Probabilistic inductive logic programming*, pages 222–243. Springer.

Raedt, Luc De, Angelika Kimmig, and Hannu Toivonen. 2007. Problog: A probabilistic prolog and its application in link discovery. In *IJCAI 2007, Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, January 6-12, 2007*, pages 2462–2467.

Richardson, Matthew and Pedro Domingos. 2006. Markov logic networks. *Mach. Learn.*, 62(1-2):107–136.

Schubert, Lenhart. 2017. Supplement for rules for probabilistic reasoning. `http:www.cs.rochester.edu/~apurtee/cogsys-supplement.pdf`.

Schubert, Lenhart K. 2004. A new characterization of probabilities in bayesian networks. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 495–503. AUAI Press.

Schubert, Lenhart K. and Chung Hee Hwang. 2000. Natural language processing and knowledge representation. chapter Episodic logic meets Little Red Riding Hood: a comprehensive natural representation for language understanding, pages 111–174. MIT Press, Cambridge, MA, USA.

Shortliffe, Edward H and Bruce G Buchanan. 1975. A model of inexact reasoning in medicine. *Mathematical biosciences*, 23(3-4):351–379.