The following full text is a publisher's version.

For additional information about this publication click this link.
http://hdl.handle.net/2066/159584

Please be advised that this information was generated on 2018-05-25 and may be subject to change.
Approximate Probabilistic Inference with Bounded Error for Hybrid Probabilistic Logic Programming

Steffen Michels¹ and Arjen Hommerson¹,² and Peter J.F. Lucas¹,³

¹Institute for Computing and Information Sciences, Radboud University, the Netherlands
²Faculty of Management, Science and Technology, Open University of the Netherlands
³LIACS, Leiden University, the Netherlands

s.michels@science.ru.nl, {arjenh,peterl} @cs.ru.nl

Abstract
Probabilistic logics, especially those based on logic programming (LP), are gaining popularity as modelling and reasoning tools, since they combine the power of logic to represent knowledge with the ability of probability theory to deal with uncertainty. In this paper, we propose a novel anytime algorithm exploiting the logical and continuous structure of distributions and experimentally show that our algorithm is, for typical relational problems, competitive with state-of-the-art sampling algorithms and outperforms them by far if rare events with deterministic structure are provided as evidence, despite the fact that it provides much stronger guarantees.

1 Introduction
Probabilistic logics are gaining popularity as modelling and reasoning tools, since they combine the power of logic to represent knowledge with the ability of probability theory to deal with uncertainty. Especially, probabilistic languages based on logic programming (LP) have become popular, as they enable knowledge representation by employing similar structure as the popular Bayesian Networks (BNs) [Pearl, 1988], but lift this from a propositional to a first-order level, thus dramatically increasing expressive power. Examples of such languages are ICL [Poole, 2008] and ProbLog [Fierens et al., 2015]. The need for such expressive languages emerged from the fact that in many areas more and more data become available, which does not only imply uncertainty, but often provides rich structure in terms of relations between entities. Probabilistic logic methods have been applied to a wide range of problem domains. Examples include link and node prediction in metabolic networks [Kimmig and Costa, 2012], dealing with a potentially unknown number of relations between multiple objects by a robot [Moldovan et al., 2012] and information fusion in the safety and security domain [Michels et al., 2013].

Probabilistic inference is an inherently computationally hard problem. However, it is well known that by exploiting logical structure as present in many distributions, exact inference is rendered feasible for many realistic problems [Chavira and Darwiche, 2008]. This is achieved by viewing the problem as a weighted model counting (WMC) problem, utilising previous work on SAT solving and model counting. These inference methods are especially suitable for probabilistic LP, as here models always have a rich logical structure (e.g. [Fierens et al., 2015]). Unfortunately, the above-mentioned structure-sensitive inference methods are usually restricted to discrete, finite distributions. Many real-world problems, however, also require continuous variables, i.e. they are best modelled using a hybrid distribution. Inference for hybrid distributions is however either made possible by significantly limiting the expressivity of the languages, e.g. [Gutmann et al., 2010; Chistikov et al., 2015], or by resorting to approximate inference methods, e.g. [Gutmann et al., 2011].

Sampling algorithms – the most widely used methods for approximate inference – are however known to converge slowly when dealing with (near) deterministic probabilities and cannot handle observed rare events well. Also inference methods based on the approximate representation of the actual continuous distribution, e.g. [Sanner and Abbasnejad, 2012; Belle et al., 2015a], suffer from observed rare events. All such methods furthermore share the drawback that they only provide weak guarantees about the quality of the estimates. In domains such as diagnosis and crisis management, where wrong decisions may have a huge impact, guarantees about the quality of approximations are highly desirable. The need to deal with rare events is also characteristic for such domains.

In this paper, we extend probabilistic LP with continuous distributions in a way that poses little restriction on the definition and use of continuous variables. Although the inference task is not computable, we show that we can provide an approximation with bounded error, which essentially provides more information about the quality of the result than the approximation methods mentioned above. The algorithm iteratively computes approximations with decreasing error, adopting a coarse-to-fine approach, making use of information from previous runs and also exploiting the logical structure of the problem to effectively decrease the error of an approximation. A comparison is made with sampling methods,
showing that in spite of the much stronger guarantees on the quality of the result, our method is competitive and even superior for hybrid problems that possess a rich logical structure and concern observed rare events.

2 Hybrid Probabilistic Logic Programming

We here introduce the language used, which is similar to Hybrid ProbLog [Gutmann et al., 2010], but allows for more expressive constraints on continuous variables.

Example 1 We consider a diagnostic problem. Suppose there are three reasons why a component i can fail: the component breaks with some probability, the temperature got too high, or a sub-component fails. Additionally, there is a cooler that cools down the temperature of all components by an amount that is uncertain as well. This cooler can also fail, indicated by the variable NoC. We can express this with the following rules:

\[
\begin{align*}
\text{fails}(i) & \leftarrow \text{Break}(i) = \text{true} \\
\text{fails}(i) & \leftarrow \text{Temp} - \text{Cooling} > \text{Limit}(i) \\
\text{fails}(i) & \leftarrow \text{Temp} > \text{Limit}(i), \text{NoC} = \text{true} \\
\text{fails}(i) & \leftarrow i \neq 0, \text{fails}(i - 1)
\end{align*}
\]

For \(i = 0\), the component does not depend on other components, so it can only fail because of the first two reasons. We can use arbitrary distributions for the continuous variables, for example:

\[
\begin{align*}
\text{Temp} & \sim N(20.0, 5.0) \\
\text{Limit}(\_ ) & \sim N(30.0, 5.0) \\
\text{Break}(\_ ) & \sim \{0.0001: \text{true}, 0.9999: \text{false}\} \\
\text{Cooling} & \sim \Gamma(30.0, 18.0) \\
\text{NoC} & \sim \{0.01: \text{true}, 0.99: \text{false}\}
\end{align*}
\]

The distributions for \text{Break}(i) and \text{Limit}(i) are assumed to be the same for all components \(i\). Assuming that a given component fails, a natural query is what is the probability that the first component also fails, e.g. \(P(\text{fails}(0)|\text{fails}(9))\).

Formally, we consider a finite number of random variables \(X, Y, Z, \ldots\), which have a support, i.e. the range of values on which the distribution is defined. These ranges are in principle arbitrary; in this paper we consider binary and real-valued random variables as typical representatives of discrete and continuous variables. For the continuous variables, we assume that their cumulative distributions functions (CDFs) can be computed exactly, thereby ignoring the potential error of typical implementations of such functions as such errors are typically negligible and predictable. We assume all random variables to be independent initially, and express dependencies by the logical structure, as common in probabilistic LP.

We use LP rules with constraints on the random variables as building blocks for the bodies, with the only restriction that satisfiability of those constraints remains decidable. Concretely, for real-numbered random variables we allow inequalities on combinations of variables. These combinations are not necessarily linear as for instance required for Conditional Linear Gaussians [Lauritzen, 1992], because constraints that consist of multiplications with real numbers are decidable. Note that this is a much weaker requirement than e.g. made for Hybrid ProbLog, which only allows a single random variable in a constraint, such that not only the constraint is decidable, but the probability of the constraint is computable as well.

As usual for probabilistic LP, the inference task we consider is the computation of the probability that a grounded predicate holds, given a conjunction of other grounded predicates.

3 Iterative Hybrid Probabilistic Model Counting

We here present a novel algorithm, the iterative hybrid probabilistic model counting (IHPMC) algorithm; it is an anytime algorithm that provides approximations with bounded error for the language outlined above.

3.1 Hybrid Probability Trees

Queries are converted to propositional formulas as is often done for probabilistic LP inference. Details of this transformation are not discussed in this paper as it is essentially the same as for ProbLog [Fierens et al., 2015]. The only difference is that the propositional formulas also include constraints on random variables. For a fixed \(i \neq 0\), the rules in Example 1 can for instance be represented as: \(\text{fails}_i \leftarrow \text{Break}, = \text{true}\lor \text{Temp} - \text{Cooling} > \text{Limit}, \lor (\text{Temp} > \text{Limit}, \lor \text{NoC} = \text{true}) \lor \text{fails}_{i-1}\).

To perform inference on such formulas we introduce the concept of Hybrid Probability Trees (HPTs). They are similar to the principles used in solvers for binary weighted model counting. Whereas in previous work the concept is applied to discrete, finite ranges, we extend the concept to continuous random variables. The basic idea is to split the variable’s range into two parts at each edge and condition the formula on the split made. To support continuous variables, we allow further splits of the same variable at deeper levels. Then, for each choice taken at a certain child, the probability is assigned that the random variable takes a value in the chosen partition, conditioned on the range chosen for the random variable earlier on the path. Such probabilities can be computed for continuous random variables by their CDFs, because we assume independence of the random variables. The splitting is repeated until the formula simplifies to \(\bot\) or \(\top\), which we for now assume to happen in a finite number of steps. By attaching a corresponding probability to each branch, this can be used to compute the probability of the event represented by the formula.

Definition 1 (Hybrid Probability Trees (HPT)) An HPT is a binary tree with at each node \(n\) a propositional formula \(\varphi_n\) and for all random variables \(X\) a range denoted by \(\text{range}(n, X)\). If \(r\) is the root node, it holds that \(\text{range}(r, X)\) equals the support of \(X\). Now let \(n\) be some node with children \(\{c_1, c_2\}\). To each edge \(n \to c_i, i \in \{1, 2\}\), a particular range \(\tau_{n_i}\) is associated to a fixed random variable \(Y\) such that \(\tau_{n_1} \cup \tau_{n_2} = \text{range}(n, Y)\) and \(\tau_{n_1} \cap \tau_{n_2} = \emptyset\). It holds that \(\text{range}(c_1, Y) = \tau_{n_1}\) and \(\text{range}(c_i, Z) = \text{range}(n, Z)\) if \(Y \neq Z\), with \(i \in \{1, 2\}\). Furthermore, the formula \(\varphi_i\) associated to \(c_1\) equals \(\varphi_n\) simplified by the restrictions on the
range on $Y$ at $c$, i.e. by observing that some primitive constraints can be replaced by $\top$ or $\bot$. Furthermore, to each edge $n \rightarrow c_i$, with $i \in \{1, 2\}$, a probability $p_{ni}$ is assigned, such that $p_{ni} = P(Y \in \tau_{ni} \mid Y \in \text{range}(n, Y))$. Finally, if $l$ is a leaf node, then it holds that $\varphi_l = \top$ or $\varphi_l = \bot$.

Given a particular HPT, it is straightforward to compute the probability of the event represented by the formula at the root node. First, the probability of a leaf is by definition $0$ or $1$. Furthermore, due to the properties of the tree it is easy to show that for each non-leaf node $n$ the probability of $\varphi_n$ given the ranges of random variables can be computed by $p_n = p_{n1} \cdot p_1 + p_{n2} \cdot p_2$, where $p_1$ and $p_2$ are the probabilities associated to $n$'s children.

Example 2 Consider a simplified diagnosis problem:

\[ \text{fail} \leftarrow T > 20.0, \ \text{NoC} = \text{true} \quad \text{fail} \leftarrow T > 30.0 \]

Here NoC represents that the cooling fails, implying a lower allowed temperature for the component. The temperature is modelled by $T \sim \mathcal{N}(20.0, 5.0)$. The event that the component is then represented by the formula $(\text{NoC} = \text{true} \land T > 20.0) \lor T > 30.0$.

A possible HPT is given in Figure 1. At each node we simplify the formula as much as possible given the choices made. Note that the probability assigned to the edge with choice $T \in (-\infty, 20]$ is not $P(T \in (-\infty, 20]) = 0.5$, as the range of $T$ is already restricted on the path before. Therefore, it is $P(T \in (-\infty, 20] \mid T \in (-\infty, 30]) \approx 0.5116$. From this tree, the probability of component failure can be computed by $0.9772 \cdot 0.01 \cdot 0.4884 + 0.0228 \approx 0.0276$.

2.3 Partially Evaluated Hybrid Probability Trees

The concept of HPT is restricted to problems for which exact inference is possible. For continuous events, which cannot be represented by hyperrectangles, we can never simplify all leaves to $\bot$ or $\top$. This problem is tackled by using partially evaluated HTs (PHPTs). For PHPTs we drop the requirement that all leaves must contain $\bot$ or $\top$ and allow arbitrary formulas. One can therefore further extend a PHPT either ad infinitum or until one finds an HPT. In the next section we show how PHPTs can be used to compute approximations with known maximal error.

Example 3 Suppose the limit of the temperature the component can bear is uncertain as well, modelled by $L \sim \mathcal{N}(30.0, 5.0)$. A PHPT for the event that the temperature is above the limit is depicted in Figure 2. It is impossible to find a finite HPT for this problem, as the event does not have the shape of a hyperrectangle. Note that the formula $T > L$ cannot be simplified directly, in case the range of only one of the variables is split. This is in contrast to binary versions of model counting algorithms, for which the choices of variable values are completely reflected in the simplified formula.

3.3 Approximating Probabilities by Partially Evaluated Hybrid Probability Trees

Bounds on Event Probabilities

Each leaf of a PHPT with a formula which is not $\bot$ or $\top$ corresponds to an area of the sample space which is only partially part of the event associated to the query. While this does not provide an exact probability of this event, it does provide a bound on this probability by assuming that the probability of such areas is $0$ or $1$ respectively. In this way, PHPTs can be used to compute probability bounds, denoted by $\underline{P}$ and $\overline{P}$.

Example 4 The PHPT in Figure 2 contains only one leaf with $\top$, whereas all other leaves contribute at most partially to the event’s probability. The lower bound of the event’s probability is therefore $\underline{P}(T > L) = 0.5 \cdot 0.9772 = 0.4886$. Analogously, as there is only one path which certainly does not contribute to the probability, the probability’s upper bound is $\overline{P}(T > L) = 1 - 0.5 \cdot 0.0228 = 0.9886$.

Proposition 1 For any event $e$ and PHPT for $e$: $\underline{P}(e) \leq \overline{P}(e)$.

All proofs are given in [Michels, 2016, Chapter 6, Appendix A]. Furthermore, we can always achieve arbitrary precision, by evaluating the PHPT sufficiently deep.

Proposition 2 For every event $e$ and maximal error $\epsilon$, there is a PHPT, such that: $P(e) - \underline{P}(e) \leq \epsilon \land \overline{P}(e) - P(e) \leq \epsilon$.

Bounds on Conditional Event Probabilities

Conditional probabilities are defined as $P(e \mid o) = P(e \land o)/P(o)$. As this definition requires the non-conditional
Input: Events e and o, maximal error $\epsilon$

Result: Approximation of $P(e \mid o)$ with max. error $\epsilon$

1. $hpt_p = PHPT$ with single root node $e \land o$
2. $hpt_n = PHPT$ with single root node $\neg e \land o$
3. $p = 0.0; \overline{p} = 1.0$
4. while $(\overline{p} - p)/2 > \epsilon$
   5. $eval_{hpt} = choose_{hpt}(hpt_p, hpt_n)$
   6. $eval_{node} = choose_{node}(eval_{hpt})$
   7. $branch_{rv} = choose_{rv}(eval_{node})$
   8. $(rv_l, rv_r) = choose_{part}(branch_{rv}, eval_{node})$
   9. add children to eval_{node} according to $(rv_l, rv_r)$
10. $p = lower(hpt_p)/(lower(hpt_p) + upper(hpt_n))$
11. $\overline{p} = upper(hpt_p)/(upper(hpt_p) + lower(hpt_n))$
12. end
13. return $(p + \overline{p})/2$

Algorithm 1: IHPMC Algorithm

As the first choice, we pick the PHPT which has the highest difference between lower and upper bound, i.e., the largest error. Second, we choose the non-evaluated node which represents the sub-space with the highest probability mass, as these nodes have the most potential for reducing the error. A more sophisticated heuristic, that tries to predict how much effort is required to reduce the error of the node’s formula, offered no improvement in experiments. The reason is that an accurate prediction takes as much time as just trying to reduce the error and continuing with another branch if the error for the first chosen one cannot be reduced quickly.

In contrast to the choice of the node, the heuristic for choosing a random variable must be a good predictor for how much this simplifies the formula, as this choice influences all descendants of the node in the tree. We use a heuristics preferring variables occurring more frequently, as also common for binary model counting [Sang et al., 2005, Section 3.2]. For continuous variables we however have to adapt the heuristic to make sure that continuous variables are not selected repeatedly without an opportunity to eliminate a primitive constraint, and thereby simplifying the formula. For instance, in Example 3, simplification of the constraint can only be achieved if both variables are selected in an alternating order.

For the binary variables there is obviously only a single partitioning possible (\{true\} and \{false\}). The partitioning of a continuous random variable amounts to picking a point, on which the range is split further. Initially, we find points that can lead to a simplified sub-formula. For instance, in Figure 2 the choice $L = (20, \infty)$ in the left branch can be used to simplify the formula to $\bot$, given the choice for $T$ above in the tree. In case there are multiple of those points, we count how many sub-formulas can be simplified by choosing this point. There are also cases in which there is no such point, as at the top or the rightmost evaluated node in Figure 2. In this case we use a heuristic trying to maximise the probability of the path eliminating the constraint.

Proposition 5 IHPMC (Algorithm 1) terminates, i.e. it always finds an approximation with the desired error bound in finite time, given the heuristics as described above.

4 Experiments

4.1 Set-up

The algorithm was implemented in the functional programming language Haskell\textsuperscript{1}. A comparison was made with the sampler implementations of BLOG 0.8\textsuperscript{2} [Milch et al., 2005] and Distributional clauses\textsuperscript{3} [Gutmann et al., 2011] (DC); both are optimised for problems with logical structure. We compare to the Likelihood weighted (BLOG LW) and the naïve rejection sampler (BLOG RJ) of BLOG; the MCMC sampler fails completely on the problem. To allow for a comparison, we ignore the bounded error guarantee of IHPMC and just consider the error of approximations. We measure the squared error and for the sampling approaches the mean

---

\textsuperscript{1}http://www.steffen-michels.de/ihpmc
\textsuperscript{2}https://bayesianlogic.github.io
\textsuperscript{3}https://github.com/davidenitti/DC
squared error over 50 runs versus inference time. All experiments were run on a Laptop with a Intel Core i3 2.4 GHz processor and 4GB of RAM. We use the following diagnostic problem:

\[
\text{fails}(i) \leftarrow \text{Break}(i) = \text{true} \\
\text{fails}(i) \leftarrow \text{Temp} > \text{Limit}(i) \\
\text{fails}(i) \leftarrow i \neq 0, \text{fails}(i-1)
\]

The number of components \( n \) and probability of \( \text{Break}(i) \) \((p)\) are varied in the experiments. For the continuous variables we use the following distributions: \( \text{Temp} \sim \mathcal{N}(20.0, 5.0) \) and \( \text{Limit}(i) \sim \mathcal{N}(\mu, 5.0) \), where \( \mu \) is a parameter that we also vary during the experiments. Note that since we use a single variable \( \text{Temp} \) for all components, dependencies are created such that probabilities cannot be computed exactly in a straightforward way.

### 4.2 Results

#### Non-Conditional Probabilities With Varying Parameters

The results depicted in Figure 3a are based on a run with parameters, where the continuous cause has a much higher probability than for a realistic diagnostic problem, but is used to illustrate behaviour of the algorithm. In the first milliseconds the error of IHPMC is very unstable. Note that the actual error depicted in the figure is not the guaranteed error bound, which is usually much larger. Actually, it takes about 2.6 seconds before the error is guaranteed to be below 0.01. Decreasing the error bound can temporarily result in an increased error, which explains that the error does not decrease monotonically. In contrast the errors of the sampling algorithms decrease more or less monotonically, as the errors were averaged over several runs.

IHPMC can profit from the case in which the discrete cause has more impact, as shown by the experiment in Figure 3b. The error bound drops below 0.01 already after about 6 milliseconds. The sampling algorithms can also profit from the discrete structure, but are clearly outperformed by IHPMC. Similarly, IHPMC can profit if the event is more rare, which is more realistic for a diagnostic problem (Figure 3c). Here it takes about 28 milliseconds to get an error bound of 0.01. In this case, IHPMC cannot make use that much of the discrete structure, but in the first iterations a single path with a high probability is found that implies \( \text{Break}_0 = \text{false} \land \text{Temp} \leq \text{Limit}_0 \land \cdots \land \text{Break}_{n-1} = \text{false} \land \text{Temp} \leq \text{Limit}_{n-1} \). This disproves \( \text{fails}_{n-1} \), which means the upper probability bound jumps to a small value.

#### Scalability

To show scalability of IHPMC we performed an experiment with a larger set of components \((n = 100)\), where it takes about 2.3 seconds to get a guaranteed error below 0.01. As the results in Figure 3d show, in the beginning two of the sampling implementations outperform IHPMC, as sampling algorithms by their nature do not suffer from a high number of dimensions. However, recall that IHPMC is outperformed by algorithms providing much less information, as no guarantees about the error are provided by the sampling algorithms. After about 1.5 second however the results of IHPMC become competitive.

### 5 Related Work

Probabilistic LP has been extended with continuous distributions, in a way that still allows for exact inference, by either restricting the events expressible [Gutmann et al., 2010] or the distributions employable [Islam et al., 2012]. There are also representations of continuous distributions that are powerful enough to approximate arbitrary distributions, but still allow for exact inference, e.g. piecewise polynomials [Sanner and Abbasnejad, 2012]. A recent algorithm for inference on such representation, which exploits the logical structure of distributions was presented in [Belle et al., 2015a]. Such approximations can in principle resemble intended distributions very closely, but there is no guarantee about the approximation error of event probabilities, as a small probability of evidence can significantly increase the error made when computing conditional probabilities.

**Distributional clauses** [Gutmann et al., 2011] is a probabilistic logic programming language, that supports arbitrary continuous distributions and events, as also supported by various other probabilistic programming languages, such as BLOG [Milch et al., 2005]. Inference for such languages is only possible by sampling approaches, usually **Markov chain Monte Carlo** (MCMC) methods, which can be very effective if they are hand-tailored for specific problems. Generic MCMC frameworks are however suboptimal and can provide poor results when rare events are observed. A significant difference with our work is that even under perfect circumstances, when detailed knowledge about the approximated distributions is available, guarantees for approximations obtained by MCMC algorithms can only be given in terms of for instance the **Monte Carlo standard error**, which gives a probabilistic, but no hard guarantee.

Recent hashing-based sampling methods, which exploit logical structure of distributions, have been extended for hybrid distributions as well [Belle et al., 2015b; Chistikov et al., 2015]. Such methods provide a probabilistic guarantee on the quality of approximations, but are restricted to exactly computable representations of continuous distributions (piecewise-polynomials [Belle et al., 2015b] and uniform distributions [Chistikov et al., 2015]). Moreover, this
remains a probabilistic guarantee on non-conditional probabilities, while IHPMC provides a hard guarantee for computed conditional probabilities.

Several approximation methods have been proposed that provide hard guarantees for discrete distributions [Poole, 1993]. Recently, very effective methods based on state-of-the-art knowledge compilation techniques have been proposed [Renkens et al., 2014; 2015; Vlasselaer et al., 2015]. In contrast to IHPMC, these algorithms are restricted to discrete problems and do not handle evidence, which are significant limitations in practice. In [Michels et al., 2015], we provided a theoretical basis for approximate inference with continuous distributions, but did not provide an efficient and practical algorithm for actually computing good approximations.

6 Conclusions

In this paper, we propose a hybrid extension for probabilistic logic programming, which allows for a much wider class of continuous distributions than existing extensions allowing for exact inference. At the same time, our extension allows one to compute approximations with bounded and arbitrarily small error. We propose a novel anytime algorithm for computing such approximations, thereby exploring the logical and continuous structure of the problem by utilising information from previous iterations. We experimentally show that our algorithm is competitive with state-of-the-art sampling algorithms, for typical relational problems, which are characterised by a moderate number of dimensions and a significant amount of structure, in spite of the fact that it provides
much stronger guarantees. Our method outperforms sampling methods by far if rare events with deterministic structure are provided as evidence. Moreover, even when our algorithm converges slowly for a certain problem, this is clearly visible to the user, thereby preventing wrong conclusions. MCMC methods do not offer similar safeguards.

References


