The following full text is a preprint version which may differ from the publisher's version.

For additional information about this publication click this link.
http://hdl.handle.net/2066/117051

Please be advised that this information was generated on 2018-01-30 and may be subject to change.
In this paper, we combine probabilistic with constraint logic programming, called probabilistic constraint logic programming (PCLP), by associating a set of constraints with a probability distribution. The idea of modeling probabilistic information using constraints was previously explored in CLP(BV) [Costa and Cussens, 2003] and an identically named language [Reizler, 1998], which both use the CLP mechanism to model a joint probability distribution on the models of a logic program. The key difference is that instead of modeling a single joint distribution, we use PCLP to define constraints on a distribution. Thus, PCLP supports imprecise probabilities, meaning that some of the probabilistic information is unknown, e.g., as in some of the earlier probabilistic logics, it is possible to specify that the probability of a fact being true is larger than 0.5.

This way of defining distributions allows PCLP to deal with continuous distributions in a powerful way. Although Hybrid ProbLog [Gutmann et al., 2010] extends ProbLog with continuously distributed facts, the logical language involving these distributions is heavily restricted, e.g., it is not possible to compare two random variables. Another approach is the work on distributional clauses, which extends the expressiveness of ProbLog further [Gutmann et al., 2011]. Although this language supports arbitrary parameterized distributions, inference is only feasible by sampling. In comparison, continuous distributions can be modeled in PCLP by providing probabilistic information about constraints in CLP(R), consistent with the actual distribution. The lower and upper bound of marginal probabilities can be derived using exact inference. The bounds can serve as approximation with desired precision, which depends on the amount of probabilistic information encoded by the constraints.

This paper is organized as follows. In the following section, PCLP is motivated by means of an example. In Sec-
3 The PCLP Language

The probabilistic constraint logic program (PCLP) language \( \mathcal{L} \) consists of a constraint language \( \mathcal{C} \), the language of rules \( \mathcal{R} \), and definitions of random variables \( \mathcal{V} \). Below, we assume the language to be fixed. Then, a PCLP \( T \) (a “theory”) consists of random variable definitions \( \mathcal{V}_T \), and a logical theory \( \mathcal{L}_T = \mathcal{C}_T \cup \mathcal{R}_T \), where \( \mathcal{C}_T \) is a constraint theory, related to the used constraints domains as discussed hereafter, and \( \mathcal{R}_T \) is a set of acyclic rules.

3.1 Syntax

Constraint Domains

In this paper, we use a language of constraints that is closed under conjunction and negation. We explore two constraint domains which are commonly used in constraint logic programming — discrete values (\( \mathbb{D} \)) and the real numbers (\( \mathbb{R} \)) — to model constraints on discrete and continuous random variables, respectively.

In the constraint domain \( \mathbb{D} \) random variables take discrete values. The constraint language consists of equality (=) and inequality (\( \neq \)) as basic building blocks. In the example we use constants such as yes in the definitions as abbreviation for the constraint that a random variable equals that constant. The constraint domain is very similar to the finite constraint domain CLP(FD) [Hentenryck, 1989], with the difference that we do not require the domain to be defined explicitly. The domain of all variables is implicitly given by all atoms occurring in the program, but one can exclude values by not assigning any probability mass to them. Using \( \mathbb{D} \), one can represent, for instance, Bayesian networks [Pearl, 1988], but also first-order formalisms such as CP-logic [Vennekens et al., 2009].

The \( \mathbb{R} \) domain is basically the same as in CLP(\( \mathbb{R} \)) [Jaffar et al., 1992]. Variables represent real numbers and constraints consist of linear equalities and inequalities using predicates such as \( =, \neq, <, >, \leq, \geq \) and functions \( \{+, -, \cdot, \div\} \). This theory can be used to approximate arbitrary continuous distributions for which the cumulative distribution function is known by associating probabilities to intervals of the domain of the distribution, defining a set of distributions including the actually intended one. To do that one can divide the probability space of a single variable in \( n \) intervals \( [l_1, u_1), \ldots, [l_n, u_n) \) and constrain its distribution as \( X \sim \{P(l_1 < X < u_1), \ldots, P(l_n < X < u_n)\} \). The number of intervals determine the precision, i.e. the gap between the probability bounds one can compute.

Rule Language

Rules are (implicitly universally quantified) expressions of the form: \( h \leftarrow l_1, \ldots, l_n \) where \( h \) is called the head and the conjunction \( l_1, \ldots, l_n \) is called the body of the rule. The head \( h \) is an atom, i.e., an expression of the form \( p(l_1, \ldots, l_m) \) with \( p \) a predicate, and \( l_1, \ldots, l_n \) are either atoms or constraints (elements of \( \mathcal{C} \)). Constants are denoted by lower-case letters \( (a, b, \ldots) \), while variables start with upper-case letters (e.g. \( X, Y, \ldots \)). As mentioned above, \( \mathcal{R}_T \) is defined as the set of rules of the PCLP theory \( T \).
Random Variable Definitions

Each random variable definition is of the form: \( V(t_1, \ldots, t_n) \sim \{ p_1 : c_1, \ldots, p_m : c_m \} \) where each \( t_i \) is a term, i.e., a constant or variable, \( c_i \in C \) is a constraint, \( p_i \in [0, 1] \) are probabilities such that \( \sum_i p_i \leq 1 \). This makes sure that random variables are defined consistently, together with the requirement that left-hand sides of definitions are non-unifiable. Given a substitution \( \theta \) that grounds \( \{ t_1, \ldots, t_n \} \), this definition yields constraints on a random variable \( V(t_1, \ldots, t_n)\theta \). The pair \( p_i : c_i \) means that the prior probability that the value of this random variable satisfies the constraint \( c_i \) is at least \( p_i \). We define the realization of a random variable, which formally captures this notion.

**Definition 1.** Let \( V \) be a random variable in a PCLP theory, where \( V \sim \{ p_1 : c_1, \ldots, p_m : c_m \} \). Then, a random variable \( X \) is called the realization of \( V \) if for all \( c_i \), \( 1 \leq i \leq m \), it holds that

\[
P(c_i[V \mapsto X]) \geq p_i
\]

where \( c[V \mapsto X] \) is the constraint \( c \) with \( V \) substituted by \( X \).

Constraints are not necessarily mutually exclusive; e.g., it is possible to express that 20% of the balls in an urn are definitely blue and 80% are either blue or yellow. However, to ensure that all the probability mass can be assigned to the domain of random variable \( V \) we require that the constraints are satisfiable. To specify the constraints on random variables, each constraint domain contains at least the associated random variables, interpreted logically as constants. To make sure random variables are defined independently, a constant in a constraint of the definition of \( V \) is either \( V \) itself or it does not unify with any variable in \( V \). For example, we may include the constraint \( \text{Support(apple)} = \text{yes} \) in the definition of \( \text{Support(banana)} \), but not refer to \( \text{Support(banana)} \).

3.2 Semantics

As the theory \( T \) specifies constraints on random variables, typically there is a class of distributions that satisfies the constraints. In fact, for a given ground atom \( q \), the theory specifies an interval \([P_{\min}(q), P_{\max}(q)]\) such that \( P_{\min}(q) \leq P(q) \leq P_{\max}(q) \).

Given a finite set of possible grounding substitutions \( \{ \theta_1, \ldots, \theta_n \} \) for each variable \( V(t_1, \ldots, t_m) \) from \( V \), we first define a probability space over a choice function \( \varphi \) which selects for each element of \( V(t_1, \ldots, t_m)\theta_i \) either one of its probability-constraint pair \( p : c_i \) or it maps to \( (1 - \sum_{(p_i : c_i) \in V} p_i : \text{true}) \) which represents the remaining probability mass. We denote the set of grounded variables by \( V_T \), which represents all the random variables in \( T \). A particular choice is then the set:

\[
CH_\varphi = \{ c \mid V \in V_T, \varphi(V) = (p : c) \}
\]

As each random variable is independent, the probability \( P_\varphi \) attached to a choice \( \varphi \) is the product of all selected probabilities:

\[
P_\varphi = \prod_{V \in V_T, \varphi(V) = (p : c)} p
\]

Each choice can be thought of as a possible world, although this world is described by constraints. In this world, it is possible that \( q \) is true, although this might depend on the actual values of the random variables involved. We say that \( s \in C \) is a solution for \( q \) given a particular choice \( \varphi \), denoted by \( \text{solution}(s, q, \varphi) \) iff

\[
L_T \cup CH_\varphi \cup \{ s \} \not\models \perp
\]

(3)

\[
L_T \cup CH_\varphi \cup \{ s \} \models q
\]

(4)

Given a particular choice function \( \varphi \), there are three possibilities: (i) the query is necessarily true in \( \varphi \), i.e., \( \text{solution(true, q, \varphi)} \) holds, denoted by \( \text{follows}(q, \varphi) \). (ii) there are no solutions or (iii) there are some constraints which imply the query, i.e., \( \exists s : \text{solution}(s, q, \varphi) \), denoted by \( \text{possible}(q, \varphi) \).

**Example 1.** Consider the following program:

\[
X \sim \{0.1: 0 \leq X \leq 1; 0.3: 1 \leq X \leq 2; 0.6: 2 \leq X \leq 3\}
\]

\[
Y \sim \{0.1: 0 \leq Y \leq 1; 0.3: 1 \leq Y \leq 2; 0.6: 2 \leq Y \leq 3\}
\]

\[
q \leftarrow Y < 0.75
\]

\[
q \leftarrow Y < 1.25, 0.25X + Y < 1.375
\]

We denote the choice selecting the \( i \)th element in the definition of \( X \) and the \( j \)th in the definition of \( Y \) as \( \varphi_{ij} \). Those choices can be represented in a two-dimensional probability space as depicted in Figure 1. Each rectangle represents an (independent) choice of a single constraint from the distribution of two random variables; the query \( q \) is represented by the area above the line. It holds that \( \text{follows}(q, \varphi_{11}) \), since for this choice \( Y < 1.25, 0.25X + Y < 1.375 \) always holds. In the grey areas, there is a possible solution, i.e., possible \((q, \varphi_{1j})\). For instance for the choice \( \varphi_{21} \), \( Y \) can be 0 and therefore less than 0.75. However, not all possible values consistent with the choice are a solution: \( Y \) can be 1 which is greater than 0.75, but less than 1.25. 0.25X + Y however can become 1.5 which is larger than 1.375.

We can use that to define the lower bound of the probability distribution, which is the probability mass of choices for which the query is certainly true. Formally, we say the lower bound of the probability for \( q \) is the sum of the probabilities of choices which imply \( q \), i.e.,

\[
P_{\min}(q) = \sum_{\text{follows}(q, \varphi)} P_\varphi
\]

(5)

For the program given in Example 1 the lower bound is \( P_{\varphi_{11}} = 0.01 \).

In contrast, in the upper bound we also include those choices which, only together with a particular solution, imply \( q \), i.e.,

\[
P_{\max}(q) = \sum_{\text{possible}(q, \varphi)} P_\varphi
\]

(6)
For the program given in Example 1 the upper bounds is $P_{\varphi_{21}} + P_{\varphi_{22}} + P_{\varphi_{12}} + P_{\varphi_{13}} = 0.22$.

We now introduce the concept of solution constraints to relate those intuitive definitions of the lower and upper bound to the possible realizations of all variables in a program.

**Definition 2.** Let a solution constraint SC$(q)$ be a constraint such that COMP$(R_T) \cup C_T \models SC(q) \rightarrow q$, where COMP$(R_T)$ is Clark’s completion [Clark, 1978] of $R_T$, i.e., rules are interpreted as if and only if statements, so that $q$ is completely characterized by SC$(q)$.

Given the solution constraint, we can now formally define the probability of $q$ given a realization of the random variables.

**Definition 3.** Given the solution constraint SC$(q)$ having random variables $V = \{V_1, \ldots, V_n\}$ and a set of independent random variables $X = \{X_1, \ldots, X_n\}$ such that $X_i$ is a realization of $V_i$, the probability of $q$, $P(q \mid X)$, is defined as $P(SC(q)[V_1 \leftarrow X_1, \ldots, V_n \leftarrow X_n])$, denoted by $P(SC(q)[V \leftarrow X])$.

The definition of bounds actually corresponds to the probability given the realizations for which the probability is minimal and maximal.

**Theorem 1.** Given the solution constraint SC$(q)$ and independent random variables $X = \{X_1, \ldots, X_n\}$ such that $X_i$ is a realization of $V_i$, where $V = \{V_1, \ldots, V_n\}$ is the set of random variables which are in SC$(q)$, it holds that $P_{\text{min}}(q)$ and $P_{\text{max}}(q)$ are the most tight bounds for $P(q \mid X)$, i.e.:

$$P_{\text{min}}(q) = \inf_X P(q \mid X) \quad P_{\text{max}}(q) = \sup_X P(q \mid X)$$

In case the constraints only relate to a single value of the random variables, then the bounds collapse, as then follows$(q, \varphi) \iff \text{possible}(q, \varphi)$. In this case, there is a simple mapping from this logic to, e.g., ProbLog.

Finally, we define negations and conjunctions of queries in a similar fashion, i.e., a conjunction of literals $q$ is defined by the constraint SC$(q)$. Note that it follows that $P_{\text{max}}(\neg q) = 1 - P_{\text{min}}(q)$ and vice versa.

We finally define what conditional probabilities mean in this setting. For each realization of $V_T$ conditional probabilities are defined in the obvious way. We then define the minimum and maximum conditional probability of a program as the probability given realizations for which this probability is minimal and maximal.

**Definition 4.** We define the probability bounds given evidence as:

$$P_{\text{min}}(q \mid e) = \inf_X \frac{P(q \wedge e \mid X)}{P(e \mid X)}$$

$$P_{\text{max}}(q \mid e) = \sup_X \frac{P(q \wedge e \mid X)}{P(e \mid X)}$$

**Example 2.** Recall the program introduced in Example 1 and consider an additional rule $e \leftarrow X < 1.5$. Suppose we would like to compute $P_{\text{max}}(q \mid e)$. To illustrate this probability, consider Figure 2, which is the probability space of Figure 1 with an additional constraint that implies the evidence. Intuitively, to determine $P_{\text{max}}(q \mid e)$ we need to exclude choices inconsistent with $e$ from the probability space and compute the probability of $q$ in the remaining space. The choices $\varphi_{31}$, $\varphi_{32}$ and $\varphi_{33}$ are certainly inconsistent with $e$ and are therefore excluded. The choices $\varphi_{21}$, $\varphi_{22}$ and $\varphi_{23}$ are possibly inconsistent with $e$, but only $\varphi_{23}$ is removed. The rationale is that, since we want to maximize the probability, removing $\varphi_{23}$ increases the resulting probability, because for all possible $X$ it could only contribute to $P(e \mid X)$ but not to $P(q \mid X)$. In contrast, removing $\varphi_{21}$ and $\varphi_{22}$ would decrease the maximal probability of $q$. In summary, we compute $P_{\text{max}}(q \mid e) = \frac{P_{\text{max}}(q \wedge e)}{P_{\text{max}}(q \wedge e) + P_{\text{max}}(\neg q \wedge e)} \approx 0.73$. Note that even though $P_{\varphi_{23}}$ would contribute to compute the upper bound of the evidence alone, it does not contribute to the partition function of $P_{\text{max}}(q \mid e)$.

To compute conditional probabilities in general we have the following proposition that relates the joint probability to the conditional probability.

**Proposition 1.** The probability bounds of a query $q$ given evidence $e$, as defined by Definition 4, can be computed as follows:

$$P_{\text{min}}(q \mid e) = \frac{P_{\text{min}}(q \wedge e)}{P_{\text{min}}(q \wedge e) + P_{\text{max}}(\neg q \wedge e)}$$

$$P_{\text{max}}(q \mid e) = \frac{P_{\text{max}}(q \wedge e)}{P_{\text{max}}(q \wedge e) + P_{\text{min}}(\neg q \wedge e)}$$

**4 Inference**

The inference problem is solved in two steps. We first apply resolution and derive a proof constraint PC$(q)$ for a query $q$, which can be seen as the operational equivalent of the solution constraint. Using this, we compute a probability for $q$ by summing over all choices. The advantage of this two-step approach is that we only need to apply resolution once. After this, we discuss strategies to reduce the number of choices we need to sum over.

**4.1 PCLP as a satisfiability problem**

The collection of a proof constraint is similar to how it is done in constraint logic programming [Jaffar et al., 1998] and ProbLog [Raedt et al., 2007]. In brief, the idea is to apply SLD resolution on the query $q$, in which the constraints encountered are collected. In the end, for each proof $i$, we...
Algorithm 1: MAXPROB

Input: proof constraint PC(q) and partial choice CHφ
Result: Pmax(q | CHφ)
1 if ¬SAT(PC(q) ∧ CHφ) then return 0
2 if ¬SAT(¬PC(q) ∧ CHφ) then return 1
3 if there is some random variable V in PC(q), for which there is no choice in CHφ then
   return \( \sum_{\varphi(V) = (p; c)} p \cdot \text{MAXPROB}(PC(q), CHφ \cup \{c\}) \)
4 else return 1

Algorithm 2: MAXPROBDÉCOMP

Input: proof constraint PC(q)
Result: Pmax(q)
1 simplify PC(q)
2 if exists independent subconstraint C in PC(q) then
   3 \( p = \text{MAXPROBDÉCOMP}(C) \)
   \( L_C = \text{MAXPROBDÉCOMP}(\text{PC}(q)[C \leftrightarrow \text{true}]) \)
   \( L_{¬C} = \text{MAXPROBDÉCOMP}(\text{PC}(q)[C \leftrightarrow \text{false}]) \)
   \( \text{return } p \cdot (L_C + (1 - p) \cdot L_{¬C}) \)
4 else return MAXPROB(PC(q), ∅)

4.2 Exact inference for PCLP

Because of the duality of \( P_{\min} \) and \( P_{\max} \), we will focus here on computing maximum probabilities. Clearly, by naively applying Lemma 1, it is necessary to consider all possible choices, which is exponential in the number of random variables. However, it is possible that a partial choice CHφ ⊆ CHφ is not consistent with or implies the proof constraint PC(q), which can be exploited during inference. To illustrate, let \( \Phi_V \) be the possible choice functions for the random variable \( V \in V_T \) only. Then, for a proof constraint such as \( X > 0 \land X + Y > 0 \), we naively examine \( |\Phi_X| \cdot |\Phi_Y| \) choices. However, assuming only half of the choices for \( X \) are consistent with \( X > 0 \), we can reduce this to only \( |\Phi_X|/2 + |\Phi_X|/2 \) examinations, if we consider the choices for \( X \) first. This effectively prunes a part of the choice space.

A basic algorithm that exploits this idea is presented in Algorithm 1. Pruning takes place at lines 1 and 2. The efficiency of the algorithm depends on the order variables are selected on line 4. As MAXPROB terminates in case an inconsistency can be found, a simple heuristic is to order random variables such that \( V_i < V_j \) if \( V_i \) occurs in an (inequality constraint with less variables on average than \( V_j \). For example, in the constraint \( X > 0 \land X + Y > 0 \) we first select choices for \( X \) as some of these choices might make the whole constraint inconsistent.

In the spirit of the well-known RelSAT algorithm [Bayardo and Pehoushek, 2000] for weighted model counting, we can also observe that in many cases the problem can be decomposed into subconstraints which do not share any random variables. For example, consider the proof constraint \( X > 0 \land X > 0 \). In this case \( P_{\max}(X > 0 \land Y > 0) = P_{\max}(X > 0) \cdot P_{\max}(Y > 0) \) which can be computed by examining \( |\Phi_X| + |\Phi_Y| \) choices only. Also for more complex queries, independent subconstraints can be found, e.g.,

\[
P_{\max}(X > 0 \land Y > 0) \land X < -3) = \]
\[
P_{\max}(X > 0) \cdot P_{\max}(X > 0) \land X < -3) + \]
\[
(1 - P_{\max}(Y > 0)) \cdot P_{\max}(X < -3)
\]

which requires to examine \( |\Phi_X| + 2 |\Phi_X| \) choices. An algorithm that uses this idea is presented in Algorithm 2. Because of the substitutions, we initially perform some basic simplifications, e.g., in the previous example \((X > 0 \land false) \land X < -3\) is simplified to \( X < -3\).

Theorem 2. Given a query atom \( q \) and its proof constraint PC(q), then

\[ \text{MAXPROBDÉCOMP}(\text{PC}(q)) = P_{\max}(q). \]
i.e., $\text{MAXPROBDECOMP}$ computes the upper bound of $q$.

5 Experiments

In this section, we provide some insight into the behavior of the proposed algorithm. In particular, we investigate the proposed heuristics and provide some experiments with respect to scalability. In the implementation we make use of the satisfiability modulo theories (SMT) solver YICES [Dutertre and Moura, 2006], which supports linear arithmetic.

5.1 Heuristics

Continuing the example of Section 2, we will first present some experiments in the computation of $P(\text{buy(apple)} \lor \text{buy(banana)} | \text{Max price(apple)} < 90)$. We compare four algorithms: (i) the naive algorithm where we sum over all choices, (ii) the pruning algorithm (Algorithm 1), (iii) an algorithm that includes the decomposition of constraints (Algorithm 1), but without pruning, and (iv) the complete algorithm proposed in this paper.

Assuming we give a point prediction as the average of $\{P_{\text{min}}(q), P_{\text{min}}(q)\}$ then the maximum error we make is $(P_{\text{max}}(q) - P_{\text{min}}(q))/2$. We varied the number of intervals into which the continuous variables were discretized and plotted the relationship between the inference time and the maximum error (cf. Figure 3a).

The result shows that the proposed heuristics improve inference time significantly. In this case, identifying the independent constraints has more effect on the inference time than the pruning heuristic, as several independences could be exploited. However, we also see that the pruning of choices reduces inference time, although the effect was limited in this case: only the choices inconsistent with $\text{Max price(apple)} < 90$ were pruned.

5.2 Scalability

To explore how scalable our inference algorithm is we add $n$ kinds of fruit to the program, for simplicity with the same characteristics as banana and determine approximations for $P(\text{buy(apple)} \lor \cdots \lor \text{buy(fruit}_n))$. The result is shown in Figure 3b where we show again the relationship between inference time and maximum error.

The result shows an interesting non-monotonic behavior: while for small $n$, the maximum error is higher if there are more fruits, the reverse holds for large $n$, e.g., if $n = 20$, then the error is smaller than for $n = 10$ given the same inference time. The first behavior occurs because the maximum error increases with the number of fruits, given the same total number of choices (see Figure 3c). So without any heuristics, the maximum error is always higher with larger $n$ given the same inference time. However, by using the heuristics the inference time for the choice space is sub-exponential, in particular because subconstraints are independent. Therefore, the size of the choice space starts to dominate and the error can even decrease given the same inference time.

6 Conclusions & Future Work

We introduced a new probabilistic constraint logic programs framework that combines CLP with probabilistic inference. Although there are already several approaches where probabilistic information is combined with logic programming, many approaches are either restricted in their capabilities to representation (e.g., only to discrete distributions) or to inference (e.g., using sampling). PCLP overcomes these issues by using a constraint representation where exact inference is feasible. In general, the language allows for specifying partially unknown probability distributions.

In this paper, we also introduced an inference algorithm exploiting state-of-the-art SMT solvers, proved its correctness with respect to the semantics, introduced several heuristics, and experimentally evaluated the implementation of the inference algorithm. We obtained encouraging results indicating that inference in our language is feasible for solving challenging problems.

In future research, we aim to improve the inference further by lifting (in [Gogate and Domingos, 2011] the closely related ReSAT algorithm is lifted) and dynamically discretizing the distribution of continuous random variables [Neil et al., 2007]. Finally, we aim to develop learning methods for this new language.

References


