The Extended Semantics For Probabilistic Programming Languages

Nicholas Hay* Vicarious

Siddharth Srivastava* UTRC Berkeley

Yi Wu UC Berkeley

Stuart Russell UC Berkeley

Abstract

Generative modeling languages, i.e., probabilistic programming languages (PPLs), offer only limited support for continuous variables; theorems concerning semantics and algorithmic correctness are often limited to discrete variables or, in some cases, bounded continuous variables. We show natural examples that violate these restrictions and break standard algorithms. Using *probability kernels*, the measure-theoretic generalization of conditional distributions, we develop the notion of *measure-theoretic Bayesian networks (MTBNs)*, and use it to provide more general semantics for PPLs with arbitrarily many random variables defined over arbitrary measure spaces. We also derive provably correct sampling algorithms for the generalized models and integrate them in the BLOG PPL.

1 Introduction

As originally defined by [Pearl, 1988], Bayesian networks express joint distributions over finite sets of random variables as products of conditional distributions. Probabilistic programming languages or PPLs [Koller *et al.*, 1997; Milch *et al.*, 2005a; Goodman *et al.*, 2008] apply the same idea to potentially infinite sets of variables. To a large extent, the semantics of these models have been defined using finite-domain, discrete variables to provide intuition, with the extension to countable domains or continuous variables left as an exercise for the reader. Moreover, it is often assumed that algorithms defined for discrete variables will work correctly with continuous variables, simply by substituting density values for discrete probability masses where needed.

For many cases, these simplifications are justified. But consider the GPA example: a two-variable Bayes net *Nationality* $\rightarrow$ *GPA*, where *Nationality* is equally likely to be India or USA. Indian GPAs range from 0 to 10, and US GPAs range from 0 to 4. Both are uniform over this range, but both have discrete probability masses, say 0.01, at the upper end because of truncation. Suppose we observe a student with a GPA of 4.0. Where does she come from?

If the student is Indian, the probability of any singleton set $\{g\}$ where $0 < g < 10$ is zero, as this range has a probability *density*, whose integral over any countable set is necessarily zero. However, if the student is American, the set $\{4\}$ has the probability 0.01. Thus, by Bayes theorem, $P(\textit{American}|\textit{GPA} \in \{4\})$ is 1. Yet if we apply the standard importance sampling algorithm[Milch *et al.*, 2005b], a sample that picks India receives a density weight of 0.99/10.0 = 0.099, whereas one that picks USA receives a discrete-mass weight of 0.01. The algorithm concludes that the student is very probably from India!

This example illustrates a fundamental problem stemming from the combination of discrete and continuous distributions, which occurs with any sensor that has thresholded limits, e.g. thermometers, weighing scales, speedometers, pressure gauges; or a hybrid sensor that can report a either real value or an error condition. Such limits result in a discrete probability for the min and max values (capturing the integration of the values that are out of range), with a continuous distribution in between. Such distributions cannot be handled by existing (even hybrid) approaches. Clearly, we could fix this by somehow counting density weights as infinitely smaller than discrete-mass weights but the situation becomes more complicated with more than one evidence variable, e.g., GPAs over multiple semesters for students who may study in both countries. Vector-valued variables also cause problems—does a point mass in three dimensions count more or less than a point mass in two dimensions? What about continuous variables with infinitely many point masses in a finite range? Existing semantics of PPLs are encumbered by several additional limitations: What about continuous random fields, which introduce uncountably many variables? What about variables that naturally require infinitely many parents, such as the time at which a Markov chain escapes a region? To cover such cases, we would ideally like a PPL to simultaneously support:

1. Random variables with infinitely (even uncountably) many parents,
2. Random variables valued in arbitrary measure spaces (with $\mathbb{R}^N$ as one case) distributed according to any measure (including discrete, continuous and mixed),
3. Establishment of conditional independencies implied by an infinite graph, and
4. Open-universe semantics in terms of the possible worlds in the vocabulary of the model.

Existing approaches do not handle most of these points. In

*The first two authors contributed equally.

particular, since they typically draw upon Kolmogorov's existence theorem [Durrett, 2013], they define the measure as a limit of a projective family over finite subsets of variables. As a consequence, they cannot assert conditional independencies involving infinitely many variables. In addition they rely on the assumption that each node has only finitely many parents to even define the projective family.

In this paper we present a measure theoretic formulation that provides all of the aforementioned desirable properties. We first define *measure-theoretic Bayesian nets (MTBNs)*, which can be used to provide semantics for any PPL. We then introduce the measure-theoretic extensions of the PPL BLOG, whose semantics provide (4) above. We prove that every well-formed BLOG model corresponds to a unique MTBN, and that every MTBN defines a unique probability measure with the properties (1-3). Finally, we present two provably correct sampling algorithms, the iterative refinement likelihood weighting (IRLW) and the lexicographic likelihood weighting (LLW), for models that include the GPA example and many others.

2 Background

We assume familiarity with measure-theoretic approaches to probability theory, but provide the fundamental definitions here. See [Durrett, 2013] and [Kallenberg, 2002] for introduction and further details. A **measurable space** $(X, \mathcal{X})$ (space, for short) is an underlying set X paired with a σ-algebra $\mathcal{X} \subseteq 2^X$ of measurable subsets of X, i.e., a family of subsets containing the underlying set X which is closed under complements and countable unions. We'll denote the measurable space simply by $\mathcal{X}$ where no ambiguity results. A function $f\colon \mathcal{X} \to \mathcal{Y}$ between measurable spaces is measurable if measurable sets pullback to measurable sets: $f^{-1}(B) \in \mathcal{X}$ for all $B \in \mathcal{Y}$. A **measure** μ on a measurable space $\mathcal{X}$ is a function $\mu\colon \mathcal{X} \to [0, \infty]$ which satisfies countable additivity: for any countable sequence $A_1, A_2, \cdots \in \mathcal{X}$ of disjoint measurable sets $\mu(\cup_i A_i) = \sum_i \mu(A_i)$. $\mathbb{P}_\mu[S]$ denotes the probability of a statement S under the base measure μ, and similarly for conditional probabilities. A probability kernel is the measure-theoretic generalization of a conditional distribution. It is commonly used to construct measures over a product space, analogously to how conditional distributions are used to define joint distributions in the chain rule.

Definition 1. *A **probability kernel** K from one measurable space $\mathcal{X}$ to another $\mathcal{Y}$ is a function $K\colon X \times \mathcal{Y} \to [0, 1]$ such that (a) for every $x \in X$, $K(x, \cdot)$ is a probability measure over $\mathcal{Y}$, and (b) for every $B \in \mathcal{Y}$, $K(\cdot, B)$ is a measurable function from $\mathcal{X}$ to $[0, 1]$.*

Given an arbitrary index set T and spaces $\mathcal{X}_t$ for each index $t \in T$, the **product space** $\mathcal{X} = \prod_{t\in T} \mathcal{X}_t$ is the space with underlying set $X = \prod_{t\in T} X_t$ the Cartesian product of the underlying sets, adorned with the smallest σ-algebra such that the projection functions $\pi_t\colon \mathcal{X} \to \mathcal{X}_t$ are measurable.

First-Order Logic We define open-universe probabilistic models using the BLOG language, whose syntax and semantics are based on typed first-order logic. Given a set of types $\mathcal{T} = \{\tau_1, \dots \tau_k\}$, a first-order vocabulary is a set of function symbols with their type signatures. Constant symbols are represented as zero-ary function symbols and predicates as Boolean functions. Given a first-order vocabulary, a possible world ("logical structure" or a "model structure") is a tuple $\langle U, I\rangle$ where the *universe* $U = \langle U_1, \dots, U_k\rangle$ and each U_i is a set of elements of type $\tau_i \in \mathcal{T}$. The *interpretation* I has, for each function symbol in the vocabulary, a function of the corresponding type signature over U.

3 Measure-Theoretic Bayesian Networks

A **digraph** G is a pair $G = (V, E)$ of a set of vertices V, of any cardinality, and a set of directed edges $E \subseteq V \times V$. Write $u \to v$ if $(u, v) \in E$, and $u \mapsto v$ if there is a path from u to v in G. A vertex $v \in V$ is a **root vertex** if there are no incoming edges to it, i.e., no $u \in V$ such that $u \to v$. Let $\mathrm{pa}(v) = \{u \in V : u \to v\}$ denote the set of parents of a vertex $v \in V$, and $\mathrm{nd}(v) = \{u \in V : \text{not } v \mapsto u\}$ denote its set of non-descendants.

A **well-founded digraph** (V, E) is one with no countably infinite ancestor chain $v_0 \leftarrow v_1 \leftarrow v_2 \leftarrow \dots$. Equivalently, one where every nonempty subset $U \subseteq V$ of vertices has at least one root vertex. This is the natural generalization of a finite directed acyclic graph to the infinite case.

Definition 2. *A **measure-theoretic Bayesian network** $M = (V, E, \{\mathcal{X}_v\}_{v\in V}, \{K_v\}_{v\in V})$ consists of (a) a well-founded digraph (V, E) of any cardinality, (b) an arbitrary measurable space $\mathcal{X}_v$ for each $v \in V$, and (c) a probability kernel K_v from $\prod_{u\in \mathrm{pa}(v)} \mathcal{X}_u$ to $\mathcal{X}_v$ for each $v \in V$.*

Fix an MTBN $M = (V, E, \{\mathcal{X}_v\}_{v\in V}, \{K_v\}_{v\in V})$. For $U \subseteq V$ let $\mathcal{X}_U = \prod_{u\in U} \mathcal{X}_u$ be the product measure space over variables $u \in U$. With this notation, K_v is a kernel from $\mathcal{X}_{\mathrm{pa}(v)}$ to $\mathcal{X}_v$. Whenever $W \subseteq U$ let $\pi^U_W\colon \mathcal{X}_U \to \mathcal{X}_W$ denote the projection map. Let $\mathcal{X}_V$ our base measure space upon which we'll consider different probability measures μ. Let X_v for $v \in V$ denote both the underlying set of $\mathcal{X}_v$ and the random variable given by the projection $\pi^V_{\{v\}}$, and X_U for $U \subseteq V$ the underlying space of X_U and the random variable given by the projection π^V_U.

Definition 3. *A MTBN M **represents** a measure μ on $\mathcal{X}_V$, if for all $v \in V$:*

- *X_v is conditionally independent of its non-descendants $X_{\mathrm{nd}(v)}$ given its parents $X_{\mathrm{pa}(v)}$.*
- *$K_v(X_{\mathrm{pa}(v)}, A) = \mathbb{P}_\mu[X_v \in A | X_{\mathrm{pa}(v)}]$ holds almost surely for any $A \in \mathcal{X}_v$, i.e., K_v is a version of the conditional distribution of X_v given its parents.*

Section 4 will show that a MTBN represents a unique measure:

Theorem 4. *A MTBN M represents a unique measure μ on $\mathcal{X}_V$.*

4 MTBNs represent unique measures

We prove here Theorem 4: A MTBN M represents a unique measure μ on $\mathcal{X}_V$. Its proof requires a series of intermediate results. We first define a projective family of measures. This gives a way to recursively construct our measure μ. We

define a notion of consistency such that every consistent projective family constructs a measure that M represents. We end by giving an explicit characterization of the unique consistent projective family, and thus of the unique measure M represents.

4.1 Consistent projective family of measures

Let K be a kernel from $\mathcal{X} \to \mathcal{Y}$ and L a kernel from $\mathcal{Y} \to \mathcal{Z}$. Their composition $K \circ L$ (note the ordering!) is a kernel from $\mathcal{X}$ to $\mathcal{Z}$ defined for $x \in X$ and $C \in \mathcal{Z}$ by:

$$(K \circ L)(x, C) = \int K(x, dy) \int L(y, dz)\, 1_C(z). \qquad (1)$$

To allow uniform notation, we will treat measurable functions and measures as special cases of kernels. A measurable function $f\colon X \to Y$ corresponds to the kernel K_f from $\mathcal{X}$ to $\mathcal{Y}$ given by $K_f(x, B) = 1(f(x) \in B)$ for $x \in X$ and $B \in \mathcal{Y}$. A measure μ on a space $\mathcal{X}$ is a kernel K_μ from 1, the one element measure space, to $\mathcal{X}$ given by $K_\mu(\cdot, A) = \mu(A)$ for $A \in \mathcal{X}$. Where this yields no confusion, we use f and μ in place of K_f and K_μ. (1) simplifies if the kernels are measures or functions. Let μ be a measure on $\mathcal{Y}_1$, K be a kernel from $\mathcal{X}_1$ to $\mathcal{Y}_1$, f be a measurable function from $\mathcal{X}_2$ to $\mathcal{X}_1$, and g be a measurable function from $\mathcal{Y}_1$ to $\mathcal{Y}_2$. Then $\mu \circ g$ is a measure on $\mathcal{Y}_2$ and $f \circ K \circ g$ is a kernel from $\mathcal{X}_2$ to $\mathcal{Y}_2$ with: $(\mu \circ g)(B) = \mu(g^{-1}(B))$, and $(f \circ K \circ g)(x, B) = K(f(x), g^{-1}(B))$.

Let Λ denote the class of upwardly closed sets: subsets of V containing all their element's parents.

Definition 5. *A **projective family** of measures is a family $\{\mu_U : U \in \Lambda\}$ consisting of a measure μ_U on $\mathcal{X}_U$ for every $U \in \Lambda$ such that whenever $W \subseteq U$ we have $\mu_W = \mu_U \circ \pi^U_W$, i.e., for all $A \in \mathcal{X}_W$, $\mu_W(A) = \mu_U((\pi^U_W)^{-1}(A))$.*

Definition 6. *Let μ be a measure on a measure space $\mathcal{X}$, and K a kernel from $\mathcal{X}$ to a measure space $\mathcal{Y}$. Then $\mu \otimes K$ is the measure on $\mathcal{X} \times \mathcal{Y}$ defined for $B \in \mathcal{X} \otimes \mathcal{Y}$ by: $(\mu \otimes K)(B) = \int \mu(dx) \int K(x, dy)\, 1_B(x, y)$.*

Definition 7. *Let K_w for $w \in W$ be kernels from $\mathcal{X}_U$ to $\mathcal{X}_{\{w\}}$. Denote by $\prod_{w \in W} K_w$ the kernel from $\mathcal{X}_U$ to $\mathcal{X}_W$ defined for each $x_U \in X_U$ by the infinite product of measures: $\left(\prod_{w \in W} K_w\right)(x_U, \cdot) = \otimes_{w \in W} K_w(x_U, \cdot)$.*

See [Kallenberg, 2002] 1.27 and 6.18 for definition and existence of infinite products of measures.

Definition 8. *A projective family $\{\mu_U : U \in \Lambda\}$ is **consistent with** M if for any $W, U \in \Lambda$ such that $W \subset U$ and $\mathrm{pa}(U) \subseteq W$, then: $\mu_U = \mu_W \otimes \prod_{u \in U \setminus W} (\pi^W_{\mathrm{pa}(u)} \circ K_u)$.*

A projective family $\{\mu_U : U \in \Lambda\}$ is consistent with M exactly when M represents μ_V:

Lemma 9. *Let μ be a measure on $\mathcal{X}_V$, and define the projective family $\{\mu_U : U \in \Lambda\}$ by $\mu_U = \pi^V_U \circ \mu$ This projective family is consistent with M iff M represents μ.*

4.2 There exists a unique consistent family

Each vertex $v \in V$ is assigned the unique minimal ordinal $d(v)$ such that $d(u) < d(v)$ whenever $(u, v) \in E$ (see [Jech, 2003] for an introduction to ordinals). For any $U \in \Lambda$ denote by $U^\alpha = \{u \in U : v(u) < \alpha\}$ the restriction of U to vertices of depth less than α. Defining $D = \sup_{v \in V}(d(v) + 1)$, the least strict upper bound on depth, we have that $U^D = U$ for all $U \in \Lambda$. In the following, fix a limit ordinal λ.

Definition 10. *$\{\nu_\alpha : \alpha < \lambda\}$ is a **projective sequence of measures** on $\mathcal{X}_{U_\alpha}$ if whenever $\alpha < \beta < \lambda$ we have $\nu_\alpha = \nu_\beta \circ \pi^{U_\beta}_{U_\alpha}$.*

Definition 11. *The limit $\lim_{\alpha<\lambda} \nu_\alpha$ of a projective sequence $\{\nu_\alpha : \alpha < \lambda\}$ of measures is the unique measure on $\mathcal{X}_U$ such that $\nu_\alpha = (\lim_{\alpha<\lambda} \nu_\alpha) \circ \pi^U_{U_\alpha}$ for all $\alpha < \beta$.*

Definition 12. *Given any $U \in \Lambda$, inductively define a measure μ^α_U on $\mathcal{X}_{U^\alpha}$ by*

$$\mu^0_U = 1,$$

$$\mu^{\alpha+1}_U = \mu^\alpha_U \otimes \prod_{v \in U : d(v) = \alpha} (\pi^{U^\alpha}_{\mathrm{pa}(v)} \circ K_v),$$

$$\mu^\lambda_U = \lim_{\alpha<\lambda} \mu^\alpha_U \qquad \text{if } \lambda \text{ is a limit ordinal.}$$

μ^α_U stabilizes for $\alpha \geq D$ to define a measure on $\mathcal{X}_U$.

The above definition is coherent as μ^α_U can be inductively shown to be a projective sequence. Theorems 13 and 14 allow us to show in Theorem 15 that $\{\mu^D_U : U \in \Lambda\}$ is the unique consistent projective family of measures. This combines with Theorem 9 to prove Theorem 4.

Theorem 13. *If $W \subseteq U$ for $W, U \in \Lambda$, then for all α: $\mu^\alpha_W = \mu^\alpha_U \circ \pi^{U^\alpha}_{W^\alpha}$.*

Theorem 14. *If $W \subset U$ where $W, U \in \Lambda$, and if $\mathrm{pa}(U) \subseteq W$, then $W^\alpha \subset U^\alpha$, $\mathrm{pa}(U^\alpha) \subseteq W^\alpha$, and $\mu^\alpha_U = \mu^\alpha_W \otimes \prod_{u \in U^\alpha \setminus W^\alpha} (\pi^{W^\alpha}_{\mathrm{pa}(u)} \circ K_u)$.*

Using the above, the following shows MTBNs satisfy properties (1-3) from the introduction:

Theorem 15. *$\{\mu^D_U : U \in \Lambda\}$ is the unique projective family of measures consistent with M.*

5 Lexicographic Likelihood Weighting (LLW)

Suppose we have a MTBN with finitely many random variables $X_1, \ldots, X_n$, and that, without loss of generality, we observe real-valued random variables $X_1, \ldots, X_m$ for $m < n$ as evidence. Suppose the distribution of X_i given its parents $X_{\mathrm{pa}(i)}$ is a mixture between a density $f_i(x_i|x_{\mathrm{pa}(i)})$ with respect to Lebesgue and a discrete distribution $F_i(x_i|x_{\mathrm{pa}(i)})$, i.e., we have $P(X_i \in [x_i - \epsilon, x_i]|X_{\mathrm{pa}(i)}) = \sum_{x \in [x_i - \epsilon, x_i]} F_i(x_i|X_{\mathrm{pa}(i)}) + \int_{x_i - \epsilon}^{x_i} f_i(x|X_{\mathrm{pa}(i)})\, dx$. Note this implies $F_i(x_i|x_{\mathrm{pa}(i)})$ is nonzero for at most countably many values x_i. If F_i is nonzero for finitely many points, it can be represented by a list of those points and their values.

Lexicographic Likelihood Weighting (LLW) extends likelihood weighting to this setting. It visits each node of the graph in topological order, sampling those variables that are not observed, and accumulating a weight for those that are observed. In particular, at an evidence variable X_i we update a tuple (d, w) of the number of densities and a weight, initially $(0, 1)$, by:

$$(d, w) \leftarrow \begin{cases} (d, wF_i(x_i|x_{\mathrm{pa}(i)})) & \text{if } F_i(x_i|x_{\mathrm{pa}(i)}) > 0, \\ (d + 1, wf_i(x_i|x_{\mathrm{pa}(i)})) & \text{otherwise.} \end{cases}$$

```
1 Type Applicant, Country;
2 distinct Country NewZealand, India, USA;
3 #Applicant(Nationality = c) {
4  if (c==USA) then ~ Poisson(50)
5  else ~ Poisson(5)};
6 origin Country Nationality(Applicant);
7 random Real GPA(Applicant s) ~
8  if Nationality(s) == USA then
9      Mix{ TruncatedNorm(3, 1, 0, 4) -> 0.9998,
10          Categorical{4 -> 1} -> 0.0001,
11          Categorical{0 -> 1} -> 0.0001}
12 else Mix{ TruncatedNorm(5, 4, 0, 10) -> 0.989,
13           Categorical{10 -> 1} -> 0.009
14           Categorical{0 -> 1} -> 0.002}};
15 random Applicant David ~
16     UniformChoice({a for Applicant a});
17 obs GPA(David) = 4;
18 query Nationality(David) = USA;
```

Figure 1: A BLOG model for the GPA example.

Finally, having produced N samples $x^{(1)}, \ldots, x^{(N)}$ by this process, let $d^* = \min_{i:w^{(i)}\neq 0} d^{(i)}$ and estimate $E[f(X)|X_{1:m}]$ by

$$\frac{\sum_{\{i:d^{(i)}=d^*\}} w^{(i)}\, f(x^{(i)})}{\sum_{\{i:d^{(i)}=d^*\}} w^{(i)}}. \tag{2}$$

Theorem 16. *LLW is consistent:* (2) *converges almost surely to* $\mathbb{E}[f(X)|X_{1:m}]$.

ue to space constraints we only sketch the proof where the evidence variables are leaves. Let x be a sample produced by the algorithm with number of densities and weight (d, w). With $I_n = \prod_{i=1,\ldots,m}(\alpha_n(x_i) - 2^{-n}, \alpha_n(x_i)]$ a 2^{-n}-cube around $x_{1:m}$ we have

$$\lim_{n\to\infty} \frac{P(X_{1:m} \in I_\epsilon | X_{m+1:n} = x_{m+1:n})}{w\, 2^{-dn}} = 1.$$

Using I_n as an approximation scheme as in the previous section, the numerator in the above limit is the weight used by IRLW. But given the above limit, using $w\, 2^{-dn}$ as the weight will give the same result in the limit. But then if we have N samples, in the limit of $n \to \infty$ only those samples $x^{(i)}$ with minimal $d^{(i)}$ will contribute to the estimation, and up to normalization they'll contribute weight $w^{(i)}$ to the estimation.

6 Preliminary Experiment Result

Fig. 1 demonstrates a BLOG program which describes a slightly modified GPA example mentioned in the previous section. It uses two types, *Applicant* and *Country*, and defines New Zealand, India and USA as distinct countries (Line 2). The number of US applicants follows a Poisson distribution with a higher mean than those from New Zealand or India (Line 3). *Origin* functions map the object being generated to the arguments that were used in the number statement that was responsible for generating it (Line 6). E.g., *Nationality* maps applicants to their nationalities. The GPA of an applicant is distributed as a mixture of weighted discrete and continuous distributions (Lines 8-14). For US applicants (lines 8-11), the range of values $0 < GPA < 4$ follows a truncated Normal$(3, 1)$ with bounds 0 and 4 (line 9). In other words, for GPAs above 4 and below 0 the density given by this component is zero (without loss of generality, we assume truncated

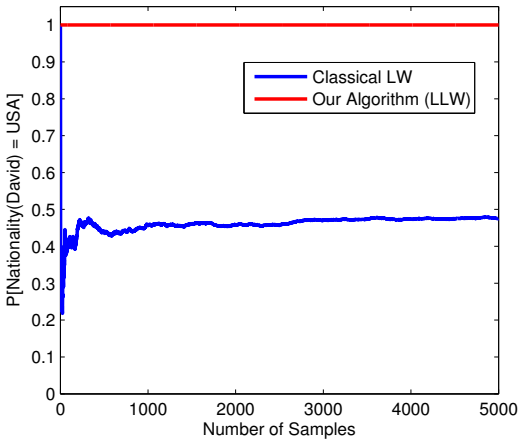

Figure 2: Estimated posterior for the GPA example in Fig. 1 using different algorithms.

distributions to be renormalized). The probability mass outside these bounds is attributed to the corresponding bounds: $P[GPA = 0] = P[GPA = 4] = 10^{-4}$. GPA distributions for other nationalities are specified similarly. Line 15-17 state that David is a random applicant with GPA equal to 4. We want to infer how likely David is from USA (Line 18).

We perform the classical likelihood weighting algorithm (classical LW) and our proposed lexicographical likelihood weighting algorithm (LLW) on this model. The estimated results for $P[Nationality(David) = USA]$ by the two algorithms are shown in Fig. 2. As we discussed above, by the Bayes rule, David must be from USA since he has a GPA of 4. Accordingly, our proposed LLW algorithm correctly computes the true posterior and converges immediately. However, for the classical LW algorithm, since it fails to handle models with this extended semantics, the estimated result is wrong.

7 Related Work

The closest related work to our framework is by Milch et al. ([2005a] [2006]), who utilize a supportive numbering of random variables, implying that each random variable has finitely many consistent parents. In addition, they only handle random variables with countably infinite ranges. The BLP framework presented by [Kersting and De Raedt, 2007] unifies logic programming with probability models, but requires each random variable to be influenced by a finite set of random variables in order to define the semantics. This amounts to requiring only finitely many ancestors of each node. Choi et al. ([2010]) present an algorithm for carrying out lifted inference over models with purely continuous random variables. They also require parfactors to be functions over finitely many random variables, thus limiting the set of influencing variables for each node to be finite. Gutmann et al. ([2011a]) also define densities over finite dimensional vectors. In a relatively more general formulation [Gutmann *et al.*, 2011b] define the distribution of each random variable using a definite clause, which corresponds to the limitation that each random variable (either discrete or continuous) has finitely many parents. Frameworks building on Markov net-

works also have similar restrictions. Wang et al. ([2008]) only consider networks of finitely many random variables, which can have either discrete or continuous distributions. Singla et al. ([2007]) extend Markov logic to infinite (non-hybrid) domains, provided that each random variable has only finitely many influencing random variables. These approaches do not address the four desirable properties discussed in the introduction.

In contrast, our approach not only allows models with arbitrarily many random variables with mixed discrete and continuous distributions, but each random variable can also have arbitrarily many parents as long as all ancestor chains are finite (but unbounded). The presented work constitutes a rigorous framework for expressing probability models with the broadest range of cardinalities (uncountably infinite parent sets) and nature of random variables (mixed, discrete, both, and even arbitrary measure spaces), with clear semantics in terms of first-order possible worlds and the generalization of conditional independences on such models. We believe that together with the foundational inference algorithms, this rigorous framework will facilitate the development of powerful techniques for probabilistic reasoning in a wide range of practical applications currently outside the scope of PPLs.

References

[Choi *et al.*, 2010] Jaesik Choi, Eyal Amir, and David J Hill. Lifted inference for relational continuous models. In *UAI*, volume 10, pages 126–134, 2010.

[Durrett, 2013] Rick Durrett. *Probability: Theory and Examples*. Cambridge University Press, 2013.

[Goodman *et al.*, 2008] Noah D Goodman, Vikash K Mansinghka, Daniel M Roy, Keith Bonawitz, and Joshua B Tenenbaum. Church: A language for generative models. In *UAI-08*, 2008.

[Gutmann *et al.*, 2011a] Bernd Gutmann, Manfred Jaeger, and Luc De Raedt. Extending problog with continuous distributions. In *Inductive Logic Programming*, pages 76–91. Springer, 2011.

[Gutmann *et al.*, 2011b] Bernd Gutmann, Ingo Thon, Angelika Kimmig, Maurice Bruynooghe, and Luc De Raedt. The magic of logical inference in probabilistic programming. *Theory and Practice of Logic Programming*, 11(4-5):663–680, 2011.

[Jech, 2003] Thomas Jech. *Set theory*. Springer, 2003.

[Kallenberg, 2002] Olav Kallenberg. *Foundations of Modern Probability*. Springer, 2002.

[Kersting and De Raedt, 2007] Kristian Kersting and Luc De Raedt. Bayesian logic programming: Theory and tool. *Statistical Relational Learning*, page 291, 2007.

[Koller *et al.*, 1997] Daphne Koller, David McAllester, and Avi Pfeffer. Effective bayesian inference for stochastic programs. In *AAAI-97*, 1997.

[Milch *et al.*, 2005a] Brian Milch, Bhaskara Marthi, Stuart J. Russell, David Sontag, Daniel L. Ong, and Andrey Kolobov. BLOG: Probabilistic models with unknown objects. In *Proc. of IJCAI*, pages 1352–1359, 2005.

[Milch *et al.*, 2005b] Brian Milch, Bhaskara Marthi, David Sontag, Stuart Russell, Daniel L. Ong, and Andrey Kolobov. Approximate inference for infinite contingent Bayesian networks. In *Tenth International Workshop on Artificial Intelligence and Statistics, Barbados*, 2005.

[Milch, 2006] Brian Christopher Milch. *Probabilistic models with unknown objects*. PhD thesis, University of California at Berkeley, Berkeley, CA, USA, 2006.

[Pearl, 1988] Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, 1988.

[Singla and Domingos, 2007] Parag Singla and Pedro Domingos. Markov logic in infinite domains. In *In Proc. UAI-07*, 2007.

[Wang and Domingos, 2008] Jue Wang and Pedro Domingos. Hybrid markov logic networks. In *AAAI*, volume 8, pages 1106–1111, 2008.