Query Algebra Operations for Interval Probabilities

Wenzhong Zhao, Alex Dekhtyar, and Judy Goldsmith

Abstract. The groundswell for the '00s is imprecise probabilities. Whether the numbers represent the probable location of a GPS device at its next sounding, the inherent uncertainty of an individual expert's probability prediction, or the range of values derived from the fusion of sensor data, probability intervals became an important way of representing uncertainty. However, until recently, there has been no robust support for storage and management of imprecise probabilities. In this paper, we define the semantics of traditional query algebra operations of selection, projection, Cartesian product and join, as well as an operation of conditionalization, specific to probabilistic databases. We provide efficient methods for computing the results of these operations and show how they conform to probability theory.

1 Introduction

Reasoning with common sense leads to attempts to make decisions and inferences with incomplete information about the real world. This process, whether applied to medical decision-making [17] or pure logical reasoning [16, 17] or to other applications such as network modeling [2] or reasoning about databases [15], can be implemented effectively using probability intervals. Models were proposed for computations with interval probabilities [4, 18, 19] recently, but there has been little work on incorporating management of interval probability distributions into databases.

Many research areas deal with discrete random variables with finite domains. Probability distributions of such random variables are finite objects, as probabilities need to be specified only for a finite number of instances. Thus, representations of probability distributions can be stored as database objects. A number of database models suitable for storing probability distributions has been proposed recently [5, 10, 13, 14].

The issue of querying *interval* probability distributions is independent of the choice of representation. As previous research on probabilistic databases has established [1, 5, 7, 13], any manipulation of probabilities in a database must be consistent with probability theory, and classical relational algebra operations fail to take this into account. Even in the case of point probabilities, one must define

the semantics of database operations carefully, as is the case in previously proposed probabilistic relational algrebras [1, 3, 7]. In order to define query algrbra operations on interval probability distributions we must define the semantics of the underlying interval probability computations. Until recently, no formal semantics for interval probabilities provided a clear and convenient way to compute marginal or conditional probability distributions. In [4], we have introduced a possible-world semantics (a generalization of a special-purpose model from [6]) for interval probability distributions which provides closed-form solutions for such computations.

This work builds upon the semantics of [4] (briefly summarized in Section 2) to introduce query algebra operations of selection, projection, conditionalization [5, 7], Cartesian product and join in databases that store interval probability distributions of discrete random variables (Section 3). The query algebra introduced in this paper is independent of any specific data model. Its operations have already been implemented in at least two (to our best knowledge) different frameworks [9, 10].

1.1 Related Work

Relational probabilistic database models were first proposed by Cavallo and Pittarelli [3] and Barbara, Garcia-Molina and Proter [1] in late-80s/early-90s. The former framework considered a single probability distribution as a complete relation; the latter used non-1NF tuples to store probability distributions. Dey and Sarkar [7] proposed a 1NF approach to storing probabilistic data and first introduced the operation of conditionalization. Kornatzky and Shimony [12] introduced the first object-oriented model for probabilistic data. All these frameworks assumed point probabilities and in all but [1] a database record/object represented information about a probability of a single event, rather than a probability distribution.

Interval probabilities were introduced to databases by Lakshmanan et. al in their ProbView [13] framework, which also used 1NF relation semantics to store probability intervals for individual events. ProbView was a predecessor of another object-oriented approach by Eiter et. al [8].

Probability distributions, rather than probabilities of individual events, became the basis of the Semistructured Probabilistic Object model introduced by Dekhtyar, Goldsmith and Hawkes [5]. In this framework, diverse discrete point probability distributions are represented as database objects. To query these objects, [5] introduced Semistructured Probabilistic Algebra. After that, Nierman and Jagadish [14] and Hung, Getoor and Subrahmnian [10] also represented probabilistic information in semistructured (XML) form.

In parallel with the development of approaches to probabilistic databases, imprecise probabilities have attracted the attention of AI researchers, as documented by the Imprecise Probability Project [11]. Walley's seminal work [18] made the case for interval probabilities as the means of representing uncertainty.

2 Semantics of Interval Probabilities

This section briefly summarizes the possible worlds semantics for interval probability distributions described in [4]. We consider the probability space $\mathcal{P} = \mathsf{C}[0,1]$, the set of all subintervals of the interval $[0,1]$. The rest of this section introduces the formal semantics for the probability distributions over $\mathcal{P}$ and the notions of *consistency* and *tightness* of interval distributions.

Definition 1. *Let V be a set of random variables. A* ***probabilistic interpretation (p-interpretation)*** *over V is a function $I_V : dom(V) \rightarrow [0,1]$, such that $\sum_{\bar{x} \in dom(V)} I_V(\bar{x}) = 1$.*

The main idea of our semantics is that a probability distribution function $P : dom(V) \rightarrow \mathsf{C}[0,1]$ represents a **set of possible point probability distributions** (a.k.a., **p-interpretations**). Given a probability distribution function $P : dom(V) \rightarrow \mathsf{C}[0,1]$, for each $\bar{x} \in dom(V)$, we write $P(\bar{x}) = [l_{\bar{x}}, u_{\bar{x}}]$. Whenever $dom(V) = \{\bar{x}_1, \ldots \bar{x}_m\}$, we write $P(\bar{x}_i) = [l_i, u_i]$, $1 \leq i \leq m$.

Definition 2. *Let V be a set of random variables and $P : dom(V) \rightarrow \mathsf{C}[0,1]$ a (possibly incomplete)*[1] *interval probability distribution function (ipdf) over V. A* probabilistic interpretation *I_V satisfies P ($I_V \models P$) iff $(\forall \bar{x} \in dom(V))(l_{\bar{x}} \leq I_V(\bar{x}) \leq u_{\bar{x}})$.*

An interval probability distribution function $P : dom(V) \rightarrow \mathsf{C}[0,1]$ is ***consistent*** *iff there exists a p-interpretation I_V such that $I_V \models P$.*

Theorem 1. *Let V be a set of random variables and $P : dom(V) \rightarrow \mathsf{C}[0,1]$ be a* complete *interval probability distribution function over V. Let $dom(V) = \{\bar{x}_1, \ldots, \bar{x}_m\}$ and $P(\bar{x}_i) = [l_i, u_i]$. P is* ***consistent*** *iff the following two conditions hold: (1) $\sum_{i=1}^{m} l_i \leq 1$; (2) $\sum_{i=1}^{m} u_i \geq 1$.*

Let $P' : X \rightarrow \mathsf{C}[0,1]$ be an incomplete *interval probability distribution function over V. Let $X = \{\bar{x}_1, \ldots, \bar{x}_m\}$ and $P'(\bar{x}_i) = [l_i, u_i]$. P' is* ***consistent*** *iff $\sum_{i=1}^{m} l_i \leq 1$.*

Consider the two interval probability distributions P and P' shown on Figure 1.(b,c) and the four p-interpretations over the same random variables from Figure 1.(a): We can see that $I_1 \models P$ and $I_1 \models P'$; $I_2 \models P$ but $I_2 \not\models P'$; $I_3 \models P'$ but $I_3 \not\models P$ and I_4 does not satisfy either P or P'.

Definition 3. *Let $P : X \rightarrow \mathsf{C}[0,1]$ be an interval probability distribution function over a set of random variables V. Let $X = \{\bar{x}_1, \ldots, \bar{x}_m\}$ and $P(\bar{x}_i) = [l_i, u_i]$. A number $\alpha \in [l_i, u_i]$ is* ***reachable*** *by P at $\bar{x}_i$ iff there exists a p-interpretation $I_V \models P$, such that $I(\bar{x}_i) = \alpha$.*

We observe that all points between any pair of reachable probability values are themselves reachable. Intuitively, points *unreachable* by an interval probability distribution do not provide any additional information about *possible* point

[1] "Incomplete", in this context means that the function need not be defined on each instance of its domain.

v v'	I_1	I_2	I_3	I_4
a a	0.3	0.45	0.3	0.25
a b	0.2	0.2	0.1	0.25
b a	0.3	0.25	0.4	0.25
b b	0.2	0.1	0.2	0.25

(a)

P:

v v'	l	u
a a	0.3	0.45
a b	0.2	0.25
b a	0.25	0.3
b b	0.1	0.25

(b)

P′:

v v'	l	u
a a	0.2	0.3
a b	0.1	0.4
b a	0.2	0.4
b b	0.1	0.2

(c)

Fig. 1. Interval Probability Distributions

probabilities. Of the possible interval probability distributions, of primary interest are those with no unreachable points. These distributions are called **tight**. We can show that for each interval probability distribution, one can find an equivalent tight probability distribution.

Definition 4. *Let $P : X \rightarrow$ C[0,1] be an interval probability distribution over a set V of random variables. P is called **tight** iff* $(\forall \bar{x} \in X)(\forall \alpha \in [l_{\bar{x}}, u_{\bar{x}}])$ (α is reachable by P at $\bar{x}$).

*Let P' be an interval probability distribution function. An interval probability distribution function P is its **tight equivalent** iff (i) P is **tight** and (ii) For each p-interpretation I, $I \models P'$ iff $I \models P$.*

Each complete interval probability distribution P has a **unique** tight equivalent. It can be efficiently computed using the following *tightening operator* $\mathcal{T}$.

Theorem 2. *Let $P : dom(V) \rightarrow$ C[0,1] be a complete interval probability distribution function over a set of random variables V. Let $dom(V) = \{\bar{x}_1, \ldots, \bar{x}_m\}$ and $P(\bar{x}_i) = [l_i, u_i]$. Then* $(\forall 1 \leq i \leq m)$

$$\mathcal{T}(P)(\bar{x}_i) = \left[\max\left(l_i, 1 - \left(\sum_{j=1}^{m} u_j - u_i \right) \right), \min\left(u_i, 1 - \left(\sum_{j=1}^{m} l_j - l_i \right) \right) \right].$$

3 Query Operations

Consider a finite collection $P_1, \ldots, P_M$ of interval probability distributions over the set of random variables V. The semantics of the operations we discuss below is independent of the representation of interval probability distributions in the database. The operations described below are applicable to any representation that has the following properties: *(i)* it is possible to retrieve/construct an individual interval probability distribution from the representation and *(ii)* there exists an efficient procedure for converting an interval probability distribution of discrete random variables into the representation. In what follows, we assume that the collection $\mathcal{D} = \{P_1, \ldots, P_M\}$ is the database.

3.1 Selection

In the relational model, selection is applied to flat tuples and its result is a collection of flat tuples. When probability distributions of discrete random variables are the database objects they can be viewed as tables in which each row describes the probability of a particular instance. Thus, selection operations on probability distributions can do two things: (i) select a subset of the input set of probability distributions and (ii) select a subset of rows in each input probability distribution. In particular, we consider three separate types of selection conditions:

Selection on participating random variables. Given a list $\mathcal{F} = \{v_1, \ldots, v_s\}$ of random variables, $\sigma_{\mathcal{F}}(\mathcal{D})$ returns the set of probability distributions that contain **all** variables in $\mathcal{F}$. The probability distributions are returned unchanged.

Selection on random variable values. Given a condition $v = x$, where $v \in \mathcal{V}$ is a random variable name and $x \in dom(v)$, $\sigma_{v=x}(\mathcal{D})$ will return the set of probability distributions which contain v as a participating random variable. In each distribution returned, *only the rows* satisfying the $v = x$ condition will remain.

Selection on probability. Given a condition $c = l \text{ op } x$ or $c = u \text{ op } x$ where x is a real number, $\text{op} \in \{=, \neq, <, >, \leq, \geq\}$ and l and u represent lower and upper bound of the probability interval, $\sigma_c(\mathcal{D})$ will return the set of probability distributions which contain at least one row satisfying the condition c. In each probability distribution returned, *only the rows* satisfying c will remain.

These operations can be illustrated on the following example. Consider the database $\mathcal{D} = \{P, P'\}$ consisting of the two probability distributions shown in Figure 1. Figure 2 shows the result of the following queries: (a) $\sigma_{\{v\}}(\mathcal{D})$ (find all distributions involving random variable v; returns both P and P' unchanged); (b)$\sigma_{v'=a}(\mathcal{D})$ (find all probabilities involving the value a of random variable v' in the database; returns two rows from each distribution); (c) $\sigma_{u=0.4}(\mathcal{D})$ (find all probability table rows with upper bound equal to 0.4; returns two rows from P').

Despite having different effects on the database, selection operations of different types commute.

Theorem 3. *Let c_1 and c_2 be two selection conditions and $\mathcal{D}$ be a database of interval probability distributions. Then $\sigma_{c_1}(\sigma_{c_2}(\mathcal{D})) = \sigma_{c_2}(\sigma_{c_1}(\mathcal{D}))$.*

3.2 Projection

As described earlier, interval probability distributions have one column per participating random variable and two additional columns for lower and upper probabilities. Here, we only consider the semantics of the projection operation that removes random variables from distributions Given a joint probability distribution of two or more random variables, the operation of obtaining a probability distribution for a proper subset of them is called in probability theory *computing the marginal probability distribution* or *marginalization*. We use the terms

$\sigma_{\{\mathbf{v}\}}(\mathcal{D})$:

v	v'	l	u
a	a	0.3	0.45
a	b	0.2	0.25
b	a	0.25	0.3
b	b	0.1	0.25

v	v'	l	u
a	a	0.2	0.3
a	b	0.1	0.4
b	a	0.2	0.4
b	b	0.1	0.2

(a)

$\sigma_{\mathbf{v}'=\mathbf{a}}(\mathcal{D})$:

v	v'	l	u
a	**a**	0.3	0.45
b	**a**	0.25	0.3

v	v'	l	u
a	**a**	0.2	0.3
b	**a**	0.2	0.4

(b)

$\sigma_{\mathbf{u}=0.4}(\mathcal{D})$:

v	v'	l	u
a	b	0.1	**0.4**
b	a	0.2	**0.4**

(c)

Fig. 2. Selection on interval probability distributions.

projection and *marginalization* interchangeably. Marginalization is a straightforward operation on point probability distribution functions [5]: given a p-interpretation I over set V of random variables and a random variable $v \in V$, the marginal probability distribution I' over set $V - \{v\}$ can be computed as follows: $\pi_{V-\{v\}}(I)(\bar{y}) = I'(\bar{y}) = \sum_{x \in dom(v)} I(\bar{y}, x)$. But when probabilities are expressed as intervals, what is a reasonable definition of the marginal probability distribution?

Recall that an interval probability distribution P is interpreted as a set of p-interpretations I satisfying it. So, when trying to represent the result of projection, we have to describe the set $\{\pi_{V-\{v\}}(I) | I \models P\}$. This intuition is captured in the following definition.

Definition 5. *Let P be an ipdf over the set V of random variables and let $U \subset V$. The result of* projection *(*marginalization*) of P onto U, denoted $\pi_U(P)$, is defined as*
$\pi_U(P)(\bar{x}) = [\min_{I \models P}(\sum_{\bar{y} \in dom(V-U)} I(\bar{x}, \bar{y})), \quad \max_{I \models P}(\sum_{\bar{y} \in dom(V-U)} I(\bar{x}, \bar{y}))]$

Definition 5 specifies precisely the result of the projection operation, but does not provide an algorithm for computing it. The following theorem presents a straightforward way to compute the projection based on a search over the space of all p-interpretations that satisfy P.

Theorem 4. *Let P be an ipdf over the set V of random variables and let $U \subset V$. Let $P'' : U \longrightarrow C[0,1]$ be*

$$P''(\bar{x}) = [\min(\sum_{\bar{y} \in dom(V-U)} l_{(\bar{x},\bar{y})}, 1), \quad \min(\sum_{\bar{y} \in dom(V-U)} u_{(\bar{x},\bar{y})}, 1)].$$

Then $\pi_U(P) = \mathcal{T}(P'')$.

The process of computing the projection $\pi_{\{v\}}(P)$ is shown below. First the random variable v' is removed from the probability distribution function. Then a new probability distribution function P'' is formed for the variable v, and the intervals $P''(a)$ and $P''(b)$ are computed as the sums, i.e. $P''(a) = [l_{(a,a)} + l_{(a,b)}, u_{(a,a)} + u_{(a,b)}]$ [2] and $P''(b) = [l_{(b,a)} + l_{(b,b)}, u_{(b,a)} + u_{(b,b)}]$. Observe that P''

[2] $l_{(x,y)}$ denotes the lower bound for $v = x$ and $v' = y$; $u_{(x,y)}$ denotes the upper bound for $v = x$ and $v' = y$.

is not tight: the upper bound of both probability intervals for $P''(a)$ and $P''(b)$ are not reachable. Thus, the final probability distribution function $\pi_{\{v\}}(P)$ is computed by applying the tightening operation on P''.

P:

v	v'	l	u
a	a	0.3	0.45
a	b	0.2	0.25
b	a	0.25	0.3
b	b	0.1	0.25

$\Longrightarrow$

v	l	u
a	0.3	0.45
a	0.2	0.25
b	0.25	0.3
b	0.1	0.25

$\Longrightarrow$

P'' :

v	l	u
a	0.5	0.7
b	0.35	0.55

$\Longrightarrow$

$\pi_v(P) = \mathcal{T}(P'')$

v	l	u
a	0.5	0.65
b	0.35	0.5

Fig. 3. Projection for interval probability distributions.

3.3 Conditionalization

Given a joint probability distribution of random variables, *selection* operation can return certain rows of its probability table and *projection* operation will compute marginal probability distributions. There is, however, another important operation on probability distributions, namely, computation of conditional probability distribution, for which no existing classical relational algebra operation seems an appropriate match. Recognizing this, Dey and Sarkar [7] proposed a new query algebra operation, *conditionalization*, which they denoted as μ.

When point probability distributions are considered, conditionalization is a straightforward operation: given a p-interpretation I over the set of random variables V, the conditional probability $\mu_{v=x}(I)$ of I under the assumption $v = x$ is computed as $\mu_{v=x}(I)(\bar{y}) = \frac{I(\bar{y},x)}{\sum_{\bar{y'} \in dom(V-\{v\})} I(\bar{y'},x)}$. When defining conditionalization on interval probability distribution functions, we follow the same intuition as with the projection operations.

Definition 6. *Let $V = (v_1, \ldots, v_n)$ be a sequence of random variables, $V^* = \{v_k, \ldots, v_n\}$, $k > 1$ and $V' = V - V^*$. Let C be a conditionalization constraint $\bar{v} = \bar{X}$, where $\bar{v} = (v_k, \ldots, v_n)$. The conditionalization is defined as*
$\mu_C(P)(\bar{y}) = \left[\min_{I \models P} \left(\frac{I_{\bar{X}}(\bar{y})}{\sum_{\bar{y}' \in dom(V')} I_{\bar{X}}(\bar{y}')}\right), \quad \max_{I \models P} \left(\frac{I_{\bar{X}}(\bar{y})}{\sum_{\bar{y}' \in dom(V')} I_{\bar{X}}(\bar{y}')}\right)\right].$

We provide a closed-form formula in [4, 9] for computing it. Notice that there are some problems inherent in this definition, as discussed in [4, 9], but that we have implemented the operation in the database with a user-beware warning.

3.4 Cartesian Product and Join

In order to consider operations that combine different interval probability distributions into one, we must first consider an issue of computing the probability of conjunctions.

Probabilistic Conjunctions. Consider two events e_1 and e_2 with known probabilities $p(e_1)$ and $p(e_2)$. When no additional information about the relationship between e_1 and e_2 is available, the probability of $e_1 \wedge e_2$ lies in the interval $[\max(0, p(e_1) + p(e_2) - 1), \min(p(e_1), p(e_2))]$. Specific assumptions about the relationship between the events may help us determine a more exact probability.

More formally, a *probabilistic conjunction* operation is a function $\otimes_\alpha : [0,1] \times [0,1] \rightarrow \mathsf{C[0,1]}$ that is *commutative*, *associative* and *monotonic* ($a \otimes_\alpha b \subseteq a \otimes_\alpha c$ iff $b \leq c$) and satisfies the following conditions: (i) $a \otimes_\alpha 0 = 0$; (ii) $a \otimes_\alpha 1 = a$; and (iii) $a \otimes_\alpha b \leq \min(a, b)$. Probabilistic conjunctions were introduced in ProbView [13] and used in other probabilistic database frameworks [6, 8]. Some examples of probabilistic conjunctions are shown in the table below:

α	$\otimes_\alpha$
independence	$a \otimes_{ind} b = [a \cdot b, a \cdot b]$
ignorance	$a \otimes_{ig} b = [\max(0, a + b - 1),\ \min(a, b)]$
positive correlation	$a \otimes_{pc} b = [\min(a, b), \min(a, b)]$
negative correlation	$a \otimes_{nc} b = [\max(0, a + b - 1),\ \max(0, a + b - 1)]$

Cartesian Product The Cartesian product of two interval probability distributions P and P' can be viewed as the joint probability distribution of the random variables from both P and P'. The resulting probability distribution will have one row $(\bar{x}, \bar{y})$ for each pair of rows $\bar{x}$ from P and $\bar{y}$ from P'. Given a relationship α together with $\otimes_\alpha$, we can define the corresponding Cartesian product of two *ipdfs*. Notice that for the Cartesian product to be defined for two distributions P and P', their sets of participating random variables V and V' *must be disjoint.*

Definition 7. *Let $P : dom(V) \rightarrow$ C[0,1] and $P' : dom(V') \rightarrow$ C[0,1] be two interval probability distributions such that $V \cap V' = \emptyset$. Let $\mathcal{I} = \{I'' : dom(V) \times dom(V') \rightarrow [0,1] | (\forall \bar{x} \in dom(V))(\forall \bar{y} \in dom(V'))(\exists I \models P)(\exists I' \models P') I''(\bar{x}, \bar{y}) \in I(\bar{x}) \otimes_\alpha I'(\bar{y})$. The* Cartesian product *$P \times_\alpha P'$ under assumption α is defined as* $(P \times_\alpha P')(\bar{x}, \bar{y}) = [\min_{I'' \in \mathcal{I}}(I''(\bar{x}, \bar{y})),\quad \max_{I'' \in \mathcal{I}}(I''(\bar{x}, \bar{y}))]$.

P:

v	v'	l	u
a	a	0.3	0.45
a	b	0.2	0.25
b	a	0.25	0.3
b	b	0.1	0.25

P$''$:

v''	l	u
a	0.5	0.6
b	0.4	0.5

P $\times_{ind}$ P$''$:

v	v'	v''	l	u
a	a	a	0.15	0.27
a	a	b	0.12	0.225
a	b	a	0.1	0.15
a	b	b	0.08	0.125
b	a	a	0.125	0.18
b	a	b	0.1	0.15
b	b	a	0.05	0.15
b	b	b	0.04	0.125

Fig. 4. Cartesian Product for interval probability distributions.

The result of Cartesian product can be computed directly.

Theorem 5. *Let $P : dom(V) \rightarrow$ C[0,1] and $P' : dom(V') \rightarrow$ C[0,1] be two interval probability distributions such that $V \cap V' = \emptyset$. Then*
$(P \times_\alpha P')(\bar{x}, \bar{y}) = [l_{\bar{x}} \otimes_\alpha l_{\bar{y}}, u_{\bar{x}} \otimes_\alpha u_{\bar{y}}]$.

Figure 4 depicts the process of computing the Cartesian product $P \times P'$ under the assumption of independence.

Join. When two *ipdfs* P and P' contain common variables, the computation of the joint distribution must ensure that the influence of common variables is accounted for only once.

Consider two interval probability distributions P and P' over sets of random variables V and V' respectively, and assume that $V \cap V' = V^*$ and $\emptyset \subset V^* \subset V \cup V'$. In what follows, let $\bar{x} \in dom(V - V^*)$, $\bar{y} \in dom(V' - V^*)$ and $\bar{z} \in dom(V^*)$.

Our goal is to define interval probability distribution $P'' : dom(V - V^*) \times dom(V^*) \times dom(V' - V^*) \rightarrow$ C[0,1], given P and P'. Consider the instance $(\bar{x}, \bar{z}, \bar{y})$ of $dom(V - V^*) \times dom(V^*) \times dom(V' - V^*)$. We can compute $P''((\bar{x}, \bar{z}, \bar{y}))$ from $P((\bar{x}, \bar{z}))$ and $P'((\bar{z}, \bar{y}))$, but with some extra effort. Direct computation of $P((\bar{x}, \bar{z})) \times_\alpha P'((\bar{z}, \bar{y}))$ is not meaningful because $\bar{z}$ affects probabilities in both P and P'. In order to be able to apply cartesian product computation, $\bar{z}$ must be factored out of one of the two distributions.

To do this, we verbalize the problem of computing $P''((\bar{x}, \bar{z}, \bar{y}))$ as "compute the joint probability of $(\bar{x}, \bar{z})$ from P and $\bar{y}$ from P', *given that V^* takes the values of $\bar{z}$*". This, in turn, suggests the use of conditionalization to factor $\bar{z}$ out of $P'((\bar{z}, \bar{y}))$. We note that in the same manner we could have attempted to factor $\bar{z}$ out of P. This leads to two families of join operations (left join and right join).

Definition 8. *Let $P : dom(V) \rightarrow$ C[0,1] and $P' : dom(V') \rightarrow$ C[0,1], and $V \cap V' = V^*$, where $\emptyset \subset V^* \subset V \cup V'$. Let α be a(n assumed) relationship between variables in V and variables in $V' - V^*$ and β be a(n assumed) relationship between variables in $V - V^*$ and V'. The operations of left and right join are defined as follows.*
$(P \ltimes_\alpha P')((\bar{x}, \bar{z}, \bar{y})) = P((\bar{x}, \bar{z})) \times_\alpha \mu_{V^* = \bar{z}}(P'((\bar{z}, \bar{y})))$.
$(P \rtimes_\beta P')((\bar{x}, \bar{z}, \bar{y})) = \mu_{V^* = \bar{z}}(P((\bar{x}, \bar{z}))) \times_\beta P((\bar{z}, \bar{y}))$.

4 Conclusions

Given the increasing interest in the use of interval probability distributions, there is a clear and present need for database methods to handle the managmenet of large collections of such data. This paper has presented the results of an initial investigation into the semantics of traditional database operations on interval probability distributions. The operations presented here are independent of any representation of the distributions in the database.

Initial implementations of these operations are being implemented by the authors of this paper as an extension of the SPO model described in [5]. This is presented in [9]. Hung, Subrahmanian and Getoor use the semantics of some of the operations presented here in their work on probabilistic XML [10].