Conditionalization for Interval Probabilities

Alex Dekhtyar
dekhtyar@cs.uky.edu
Department of Computer Science
University of Kentucky

Judy Goldsmith
goldsmit@cs.uky.edu
Department of Computer Science
University of Kentucky

January 21, 2002

Abstract

Conditionalization, i.e., computation of a conditional probability distribution given a joint probability distribution of two or more random variables is an important operation in some probabilistic database models. While the computation of the conditional probability distribution is straightforward when the exact point probabilities are involved, it is often the case that such exact point probability distributions of random variables are not known, but are known to lie in a particular interval.

This paper investigates the conditionalization operation for interval probability distribution functions under a possible world semantics. In particular, given a joint probability distribution of two or more random variables, where the probability of each outcome is represented as an interval, we (i) provide formal model-theoretic semantics; (ii) define the operation of conditionalization and (iii) provide a closed form solution/efficient algorithm to compute the conditional probability distribution.

subarea and keywords: conditionals in probabilistics and possibilistics; probability intervals, probabilistic databases, conditionalization.

1 Introduction

The problem of finding conditional probability given a joint probability of two events is as old as probability theory itself. The famous Bayes formula,

$$P(A|B) = P(AB)/P(B),$$

lies at the core of Bayesian probability theory. It is no surprise, therefore, that database models for managing probabilistic data include the operation of *conditionalization*, the procedure for obtaining the conditional probability distribution. The first occurrence of conditionalization in a probabilistic relational algebra is due to Dey and Sarkar [4]. Conditionalization operation is also used by Dekhtyar, Goldsmith and Hawkes in their Semistructured Probabilistic Database model [2]. In both of these models, the exact probabilities are assumed to be known, so it is possible to use the Bayes formula to compute the conditional probability.

In many situations, however, exact probabilities of events/ exact probability distributions are not available, but interval estimates for each probability can be obtained. Such situations include, but are not limited to, domains where (i) probabilities are derived via limited sampling, or (ii) probabilities are elicited from experts as intervals, or (iii) probabilities are obtained as the combination of dissimilar sets of probabilities.

Work on imprecise probabilities, and interval probability in particular, falls into two categories: mathematical foundations and use in applications. We focus here on the former. Our work differs significantly from the major foundational work so far [10, 11, 8] because we consider a different semantics (as compared to Walley, for instance [10]) or the same semantics applied to different objects (see Weichselberg [11]). The semantics we consider can be described as a *possible worlds* semantics.

If probabilities are elicited as *gambles*, as described by Walley [10], the behavioural semantics [8] leads to a computation of conditionalization of upper and lower previsions separately and independently, rather than direct conditionalization of probability intervals. On the other hand, Weichselberger [11] gives a possible world semantics for probability intervals. However, his semantics, and therefore his definition of conditional probability, apply to Kolmogorov-style probability structures based on atomic events, as opposed to join interval probability distributions of discrete random variavbles.

In this work we solve the following problem. Consider a set of random variables and a joint probability distribution of two or more of them, such that the probability of each outcome is expressed as a subinterval of the interval $[0, 1]$. Given a condition on one of the random variables participating in this joint distribution, compute the resulting conditional probability distribution. More formally, if $v_1, \ldots, v_n$ are random variables, $P(v_1, \ldots, v_n)$ is an *interval probability distribution function*, and $x \in dom(v_n)$, we want to find a function $P'(v_1, \ldots, v_{n-1})$ that $P'(v_1 = a_1, \ldots, v_{n-1} = a_{n-1}) = P(v_1 = a_1, \ldots, v_{n-1} = a_{n-1} | v_n = x)$.

In order to adress this problem, we first need some defintions. In Section 2 we define *interval probability distribution functions*. In this section we also describe the semantics of these functions and their key properties. Assuming this semantics, we describe the conditionalization problem in Section 3. In particular, we define the conditionalization operator on interval probability distribution functions and then provide an efficient mechanism for computing this operator. The two major contributions of this paper are

- Model-theoretic semantics for interval probability distribution functions, and
- Closed-form solutions for conditionalization of interval probability distribution functions and efficient algorithms for computing conditionalization *for the proposed semantics for interval probability distribution functions*.

It is important to note that the meaning of the conditionalization problem, and therefore its solution, depend greatly on one's interpretation of what an *interval probability distribution function* is. While we feel that the model presented here is a reasonable interpretation, it is not the only possible one. Our results on conditionalization hold only with respect to the possible world semantics of the interval probability distribution functions.

2 Semantics of Interval Probability Distributions

2.1 Consistency and p-interpretations

In this paper we assume that the probability space $\mathcal{P} = \mathsf{C[0,1]}$ is the set of all subintervals of the interval $[0, 1]$. The rest of this section introduces the notions of formal semantics for the probability distributions over $\mathcal{P}$ and the notions of *consistency* and *tightness* of the distributions. Similar treatment of interval probability distributions appeared for the first time in [3], where interval probability distributions were discussed in the context of Temporal Probabilistic Databases. Here, we give a more general version of the semantic framework and extend it to match our goals.

We consider a finite universe $\mathcal{V}$ of discrete random variables $v_1, v_2, \ldots v_N$. For each random variable $v \in \mathcal{V}$, $dom(v)$ denotes the set of possible values v can take. If $V = (v_1, \ldots, v_k)$ is a sequence of random variables, then $dom(V)$ denotes $dom(v_1) \times dom(v_2) \times \ldots \times dom(v_k)$.

Definition 1 *Let $V = (v_1, v_2, \ldots, v_k)$ be a sequence of random variables from (V). An* interval probability distribution function[1] *over V is any function $P : dom(V) \rightarrow$ C[0,1].*

Where the use of it does not cause confusion, we abbreviate "probability distribution function" to "pdf" and "interval probability distribution function" to "ipdf".

There is one major omission in the definition above. Remember that a (complete) *point* probability distribution over V is defined as a function $p : dom(V) \rightarrow [0, 1]$, *such that* $\sum_{\bar{x} \in dom(V)} p(\bar{x}) = 1$. The latter condition specifies which of the functions $f : dom(V) \rightarrow [0, 1]$ can be considered *valid* probability distribution functions. In what follows, we investigate such a validity criterion for the ipdfs.

Our first goal is to interpret the interval as the probability of a particular outcome. We approach this problem by assuming that in the "real world" the probability of any outcome is a *point probability*. This means that the "real" probability distribution function for the joint distribution of random variables V in our world is a *point probability distribution*. The intervals represent our *lack of knowledge* about the exact point probability distribution. Therefore, we assume that an interval probability distribution represents a set of *possible* point probability distributions.

Definition 2 *Let V be a sequence of random variables. A* **probabilistic interpretation (p-interpretation)** *over V is a function $I_V : dom(V) \rightarrow [0, 1]$, such that $\sum_{\bar{x} \in dom(V)} I_V(\bar{x}) = 1$.*

Given a set of random variables, any valid point probability distribution is a *p-interpretation* over it. Given an ipdf, a p-interpretation plays the role of a "possible point probability distribution" as mentioned above.

In the rest of the paper we adopt the following notation. Given a probability distribution funtion P : $dom(V) \rightarrow$ C[0,1], we write for each $\bar{x} \in dom(V)$, $P(\bar{x}) = [l_{\bar{x}}, u_{\bar{x}}]$. Whenever we enumerate $dom(V)$ as $dom(V) = \{\bar{x}_1, \ldots \bar{x}_m\}$, we write $P(\bar{x}_i) = [l_i, u_i]$, $1 \leq i \leq m$. Also, it is sometimes convenient to write $I = (a_1, a_2, \ldots, a_m)$ to represent a p-interpretation I such that $I(\bar{x}_i) = a_i$, $1 \leq i \leq m$.

Definition 3 *Let V be a set of random variables and $P : dom(V) \rightarrow$ C[0,1] an ipdf over V. A* probabilistic interpretation *I_V satisfies P ($I_V \models P$) iff*

$$(\forall \bar{x} \in dom(V))(l_{\bar{x}} \leq I_V(\bar{x}) \leq u_{\bar{x}}).$$

Basically, if a p-interpretation I_V satisfies an interval pdf P, then given P, I_V is a *possible point probability distribution*. As such, we interpret an ipdf P as a set $\{I_V | I_V \models P\}$. These I_Vs are the "possible worlds" represented by the ipdf.

Example 1 *Consider a random variable v with domain $\{a, b, c\}$. Let probability distribution functions P_1, P_2 and P_3 and p-interpretations I_1, I_2, I_3 and I_4 be defined as*

P_1	P_2	P_3	I_1	I_2	I_3	I_4
$P_1(a) = [0.2, 0.3]$	$P_2(a) = [0.3, 0.6]$	$P_3(a) = [0.4, 0.5]$	$I_1(a) = 0.3$	$I_2(a) = 0.5$	$I_3(a) = 0.25$	$I_4(a) = 0.7$
$P_1(b) = [0.3, 0.45]$	$P_2(b) = [0.3, 0.4]$	$P_3(b) = [0.4, 0.5]$	$I_1(b) = 0.3$	$I_2(b) = 0.4$	$I_3(b) = 0.45$	$I_4(b) = 0.3$
$P_1(c) = [0.3, 0.5]$	$P_1(c) = [0, 0.4]$	$P_3(c) = [0.4, 0.5]$	$I_1(c) = 0.4$	$I_2(c) = 0.1$	$I_3(c) = 0.3$	$I_4(c) = 0$

[1]Here, we consider only *complete* interval probability distribution functions, i.e., functions which provide probability estimates for *each* possible combination of values of the participating random variables. The framework described here can be extended to the case of incomplete probability distributions.

P-interpretation I_1 satisfies both P_1 and P_2. P-interpretation I_2 satisfies P_2 but not P_1 while I_3 satisfies P_1 but not P_2. Finally, I_4 satisfies neither P_1 nor P_2. None of the p-interpretations I_1, I_2, I_3, I_4 satisfies P_3.

Now consider an arbitrary ipdf P. As mentioned above, Definition 1 lacks a validity criterion for ipdfs. We can reconstruct this criterion now.[2]

Definition 4 *An interval probability distribution function $P : dom(V) \to$ C[0,1] is **consistent** iff there exists a p-interpretation I_V, such that $I_V \models P$.*

From now on, we consider only consistent ipdfs. This excludes all ipdfs which have no satisfying p-interpretations. Example 2 below illustrates this definition.

Example 2 *Consider the ipdfs P_1, P_2 and P_3 described in Example 1. As that example shows $I_1 \models P_1$ and $I_1 \models P_2$, so both P_1 and P_2 are* consistent *ipdfs.*

On the other hand, none of the p-interpretations from Example 1 satisfied P_3. Any p-interpretation I satisfying P_3 must have $I(a) \geq 0.4$, $I(b) \geq 0.4$ and $I(c) \geq 0.4$, hence $I(a) + I(b) + I(c) \geq 1.2$, which contradicts the constraint $I(a) + I(b) + I(c) = 1$ on p-interpretations. No p-interpretation could satisfy P_3; P_3 is inconsistent.

Given an interval pdf P, there is a straightforward procedure that allows one to check P's consistency. The procedure is based on the following result.

Theorem 1 *Let V be a set of random variables and $P : dom(V) \to$ C[0,1] be an ipdf 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 conditions hold: (i) $\sum_{i=1}^{m} l_i \leq 1$ and (ii) $\sum_{i=1}^{m} u_i \geq 1$.*

Proof.

Let P be an ipdf over V and let $dom(V) = \{\bar{x}_1, \ldots, \bar{x}_m\}$. Remember that we denote $P(\bar{x}_i)$ as $[l_i, u_i]$. Consider now two functions $f_l, f_u : dom(V) \to [0,1]$ such that $f_l(\bar{x}_i) = l_i$ and $f_u(\bar{x}_i) = u_i$ for all $1 \leq i \leq m$.

1. $\underline{\sum_{i=1}^{m} l_i \leq 1 \text{ and } \sum_{i=1}^{m} u_i \geq 1 \Rightarrow P \text{ is consistent.}}$

If $\sum_{i=1}^{m} l_i = 1$ then f_l is a p-interpretation and $f_l \models P$, therefore P is consistent.

If $\sum_{i=1}^{m} u_i = 1$ then f_u is a p-interpretation and $f_u \models P$, therefore P is consistent.

Consider now the case when $\sum_{i=1}^{m} l_i < 1$ and $\sum_{i=1}^{m} u_i < 1$. Let $\sum_{i=1}^{m} l_i = L$ and $\sum_{i=1}^{m} u_i = U$. We know that $L < 1 < U$.

Consider a function $I : dom(V) \to [0,1]$ such that

$$I(\bar{x}_i) = \frac{1-L}{U-L} u_i + \left(1 - \frac{1-L}{U-L}\right) l_i.$$

We now show that I is a p-interpretation and $I \models P$. Let $\alpha = \frac{1-L}{U-L}$. As $L < 1 < U$, $0 < \alpha < 1$ and we can rewrite the definition of I as $I(\bar{x}_i) = l_i + \alpha(u_i - l_i)$. Then $l_i \leq I(\bar{x}_i) \leq u_i$. Thus, if I is a p-interpretation then $I \models P$.

To show that I is a p-interpretation we need $\sum_{i=1}^{m} I(\bar{x}_i) = 1$. This can be demonstrated as follows:

$\sum_{i=1}^{m} I(\bar{x}_i) = \sum_{i=1}^{m} (\frac{1-L}{U-L} u_i + (1 - \frac{1-L}{U-L}) l_i) = \alpha \sum_{i=1}^{m} u_i + (1-\alpha) \sum_{i=1}^{m} l_i = \alpha U + (1-\alpha) L = \frac{1-L}{U-L} U + (1 - \frac{1-L}{U-L}) L = \frac{(1-L)U + (U-L-1+L)L}{U-L} = \frac{U - LU + LU - L^2 - L + L^2}{U-L} = \frac{U-L}{U-L} = 1$.

[2]For historical reasons we refer to this criterion as "consistency".

X	l	u
$\bar{x}_1$	0.1	0.2
$\bar{x}_2$	0.1	0.2
$\bar{x}_3$	0.1	0.3
$\bar{x}_4$	0.1	0.8

X	l	u
$\bar{x}_1$	0.1	0.2
$\bar{x}_2$	0.1	0.2
$\bar{x}_3$	0.1	0.3
$\bar{x}_4$	0.3	0.7

Figure 1: Tightness of interval probability distributions.

2. $\sum_{i=1}^{m} l_i \leq 1$ and $\sum_{i=1}^{m} u_i \geq 1 \Leftarrow P$ is consistent.

If P is consistent, then there exists a p-interpretation $I : dom(V) \rightarrow [0,1]$, such that $I \models P$. Then, for each $\bar{x}_i, 1 \leq i \leq m$, we have $l_i \leq I(\bar{x}_i) \leq u_i$. But then,

$\sum_{i=1}^{m} l_i \leq \sum_{i=1}^{m} I(\bar{x}_i) \leq \sum_{i=1}^{m} u_i$.

As I is a p-interpretation, $\sum_{i=1}^{m} I(\bar{x}_i) = 1$ and we immediately get $\sum_{i=1}^{m} l_i \leq 1$ and $\sum_{i=1}^{m} u_i \geq 1$. □

2.2 Tightness

Consistency is not the only property of interval probability distribution functions of interest. Another property, *tightness*, is also very important. A similar notion arises in Walley's work [10] (he talks about "avoiding sure loss"). In Weichselberger [11], tightness shows up as *F-probability*.

Example 3 *Consider the ipdf P as shown in Figure 1 (left). It is easy to see that P is consistent (indeed, the sum of lower bounds of probability intervals adds up to 0.4 and the the sum of the upper bounds adds up to 1.5). In fact, there will be many different p-interpretations satifying P. Of particular interest to us are the p-interpretations that satisfy P and take on marginal values. E.g., p-interpretation I_1: $I_1(\bar{x}_1) = 0.1; I_1(\bar{x}_2) = 0.1; I_1(\bar{x}_3) = 0.1; I_1(\bar{x}_4) = 0.7$ satisfies P and hits the lower bounds of probability intervals provided by P for $\bar{x}_1$, $\bar{x}_2$ and $\bar{x}_3$. Similarly, I_2: $I_2(\bar{x}_1) = 0.2; I_2(\bar{x}_2) = 0.2; I_2(\bar{x}_3) = 0.3; I_2(\bar{x}_4) = 0.3$ satisfies P and hits the upper bounds of probability intervals for $\bar{x}_1, \bar{x}_2$ and $\bar{x}_3$. Thus, every single number in the probability intervals for $\bar{x}_1$, $\bar{x}_2$ and $\bar{x}_3$ is reachable by different p-interpretations satisfying P.*

However, the same is not true for $\bar{x}_4$. It is easy to see that for no p-interpretation I satisfying P, $I(\bar{x}_4) = 0.1$. Indeed, we know that $I(\bar{x}_1)+I(\bar{x}_2)+I(\bar{x}_3)+I(\bar{x}_4) = 1$ and if $I(\bar{x}_4) = 0.1$ then $I(\bar{x}_1)+I(\bar{x}_2)+I(\bar{x}_3) = 0.9$. However, the maximum values for $\bar{x}_1$, $\bar{x}_2$ and $\bar{x}_3$ allowed by P are 0.2, 0.2 and 0.3 respectively, and they add up to only 0.7.

Similarly, no p-interpretation I satisfying P can have $I(\bar{x}_4) = 0.8$. Indeed, in this case, $I(\bar{x}_1) + I(\bar{x}_2) + I(\bar{x}_3) = 1 - 0.8 = 0.2$. However, the smallest values for $\bar{x}_1$, $\bar{x}_2$ and $\bar{x}_3$ allowed by P are all 0.1 and they add up to 0.3.

The "reachability" notion discussed in the example above can be formalized.

Definition 5 *Let $P : X \rightarrow$ 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 $I(\bar{x}_i) = \alpha$.*

The proposition below specifies the key property of reachability: it is continuous, i.e., every point between two reachable points is also reachable.

Proposition 1 *Let $P : X \rightarrow$ C[0,1] be an interval probability distribution function over a set of random variables V. If for some $\bar{x} \in X$ there exist α, β, $l_{\bar{x}} \leq \alpha \leq \beta \leq u_{\bar{x}}$ which are both reachable by P at $\bar{x}$, then* ***any*** *$\gamma \in [\alpha, \beta]$ is reachable by P at $\bar{x}$.*

Proof. Similar to the proof of Theorem 1. □

Intuitively, points *unreachable* by an ipdf represent "dead weight"; they do not provide any additional information about the *possible* point probabilities. Dealing with such interval pdfs is inconvenient - unreachable points obscure the actual structure of the set of satisfying p-interpretations. It is desirable to consider only those interval pdfs that contain no unreachable points. We need, however to make sure that (i) no expressive power is lost by ignoring ipdfs with unreachable points and (ii) given an ipdf with unreachable points there is a way to "deflate" it, i.e., to identify and "remove" all unreachable points.

Definition 6 *Let $P : X \rightarrow$ C[0,1] be an ipdf 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 $\bar{x}$).

Example 4 *Let us pick up where Example 3 left off. As shown in that example, the ipdf P shown on the left-hand side of Figure 1 is* not tight. *On the other hand, the ipdf P' on the right-hand side of Figure 1 is* tight. *Its tightness follows from the fact that p-interpretations I_1 and I_2 from Example 3 both satisfy it, and now, both upper and lower bounds for $\bar{x}_4$ are reachable. By Proposition 1 this means that every point between upper and lower bound is reachable for $\bar{x}_4$.*

Function P' has another important distinction w.r.t. to P. Indeed, one can show that for any p-interpretation I, $I \models P$ iff $I \models P'$, i.e., the sets of p-interpretations that satisfy P and P' coincide. Hence, one can say that P' is a tight equivalent *of P.*

In general when dealing with interval probability distributions that are not tight, we will want to replace them with their *tight equivalents*. The procedure of substituting an untight interval pdf with its tight equivalent, which we call *tightening* is exactly the "deflation" mechanism that we mentioned above.

Definition 7 *Given an ipdf P, an ipdf P' is its* ***tight equivalent*** *iff*

- *P' is* ***tight****.*
- *For each p-interpretation I, $I \models P$ iff $I \models P'$.*

An ipdf P uniquely determines the set of p-interpretations that satisfy it. Hence,

Proposition 2 *Each interval probability distribution function P has a* ***unique*** tight equivalent.

Definition 8 *A* ***tightening*** *operator $\mathcal{T}$ takes as input an interval probability function $P : X \rightarrow$ C[0,1] and returns its* ***tight equivalent*** *$P' : X \rightarrow$ C[0,1].*

Our next goal is to compute the result of applying the tightening operator to an interval probability distribution function efficiently. First we notice that if P is tight then $\mathcal{T}(P) = P$. The intuition behind tightening can be shown in the following example.

Example 5 *We continue studying the interval pdfs in Figure 1. As shown in Examples 3 and 4, the ipdf* P *on the left is not tight, and the ipdf* P' *on the right is its tight equivalent. But how, given* P*, do we construct* $P' = \mathcal{T}(P)$ *?*

We know that P *is not tight only for* $\bar{x}_4$*, so from this we can conclude that* $\mathcal{T}(P)(\bar{x}_i) = P(x_i)$ *for* $i = 1, 2, 3$*. Consider now* $\bar{x}_4$*. What is the largest possible probability that* $\bar{x}_4$ *can have? Well, this probability is maximized when the probabilities of all other outcomes are minimized. This occurs when the probabilities take the value of their respective lower bounds. In our case, we can minimize the sum of probabilities for outcomes* $\bar{x}_1, \bar{x}_2$ *and* $\bar{x}_3$ *at* $0.3 = 0.1 + 0.1 + 0.1 = l_1 + l_2 + l_3$*. As the sum of all probabilities must be equal to 1, the probability of* $\bar{x}_4$ *must take the value of* $1 - (l_1 + l_2 + l_3) = 1 - 0.3 = 0.7$*. Since* $0.7 \in [l_4, u_4] = [0.1, 0.8]$*, the p-interpretation* I *defined as* $(0.1, 0.1, 0.1, 0.7)$ *satisfies* P*.*

It is now clear that no p-interpretation I' *such that* $I'(\bar{x}_4) > 0.7$ *can satisfy* P*: constraints* $0.1 \le I'(\bar{x}_1)$*,* $0.1 \le I'(\bar{x}_2)$*,* $0.1 \le I'(\bar{x}_3)$*,* $0.7 < I(\bar{x}_4)$*, and* $I'(\bar{x}_1) + I'(\bar{x}_1) + I'(\bar{x}_1) + I'(\bar{x}_1) = 1$ *cannot be simultaneously satisfied.*

Similar reasoning about the smallest possible probability value for $\bar{x}_4$ *yields* $0.3 = 1-(0.2+0.2+0.3) = 1 - (u_1 + u_2 + u_3)$ *as the answer. P-interpretation* $I'' = (0.2, 0.2, 0.3, 0.3)$ *satisfies* P *and for any p-interpretation* I^* *such that* $I^*(\bar{x}_4) < 0.3$*, constraints* $I^*(\bar{x}_1) \le 0.2$*,* $I^*(\bar{x}_2) \le 0.2$*,* $I^*(\bar{x}_3) \le 0.3$*,* $I^*(\bar{x}_4) < 0.3$*, and* $I^*(\bar{x}_1) + I^*(\bar{x}_1) + I^*(\bar{x}_1) + I^*(\bar{x}_1) = 1$ *cannot be satisfied. Combining these two observations together, we conclude that* $\mathcal{T}(P)(\bar{x}_4) = [0.3, 0.7] = [1 - (u_1 + u_2 + u_3), 1 - (l_1 + l_2 + l_3)]$*.*

In practice, there are other cases to consider. Theorem 2 specifies the exact closed form solution for the tightening operator. This solution induces an efficient procedure for computing the results of tightening an interval probability distribution function.

Theorem 2 *Let* $P : dom(V) \to$ *C[0,1] be a consistent ipdf 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* $(\forall i)(1 \le i \le m)$

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

Proof(*sketch*)

Let $P'(\bar{x}_i) = [\max(l_i, 1 - \sum_{j=1}^{m} u_j + u_i), \min(u_i, 1 - \sum_{j=1}^{m} l_j + l_j)]$. We need to prove two statements: $P \equiv P'$ and P' is tight.

• $\underline{P \equiv P'}$.

First we notice that for all $1 \le i \le m$, $[\max(l_i, 1 - \sum_{j=1}^{m} u_j + u_i), \min(u_i, 1 - \sum_{j=1}^{m} l_j + l_j)]. \subseteq [l_i, u_i]$.

Indeed, $l_i \le \max(l_i, 1 - \sum_{j=1}^{m} u_j + u_i)$ and $\min(u_i, 1 - \sum_{j=1}^{m} l_j + l_j)] \subseteq [l_i, u_i] \le u_1$. Now, because P is consistent, $\forall 1 \le i \le m$, $l_i \le u_i$ and $\sum_{j=1}^{m} l_j \le 1 \le \sum_{j=1}^{m} u_j$. But then, $1 - \sum_{j=1}^{m} u_j \le 0$ and hence $1 - \sum_{j=1}^{m} u_j + u_i \le u_i$, and therefore $\max(l_i, 1 - \sum_{j=1}^{m} u_j + u_i) \le u_i$.

Similarly, we obtain $l_i \le \min(u_i, 1 - \sum_{j=1}^{m} l_j + l_j)]$. Finally, for $1 \le j \le m$, $l_j \le u_j$, $\sum_{j=1}^{m} l_j - l_i \le \sum_{j=1}^{m} u_j - u_i$ and therefore $1 - \sum_{j=1}^{m} u_j + u_i \le 1 - \sum_{j=1}^{m} l_j + l_i$. Therefore,

$$l_i \le \max(l_i, 1 - \sum_{j=1}^{m} u_j + u_i) \le \min(u_i, 1 - \sum_{j=1}^{m} l_j + l_j) \le u_i.$$

This means that $(\forall I : dom(V) \to [0, 1])(I \models P' \Rightarrow I \models P)$.

We now need to show the inverse: $(\forall I : dom(V) \to [0, 1])(I \models P \Rightarrow I \models P')$.

Let I be a p-interpretation over V and let $I \models P$. Therefore, $(\forall 1 \leq i \leq m)(l_i \leq I(\bar{x}_i) \leq u_i)$. We need to show $\max(l_i, 1 - \sum_{j=1}^m u_j + u_i) \leq I(\bar{x}_i) \leq \min(u_i, 1 - \sum_{j=1}^m l_j - l_j)$.

We show $\max(l_i, 1 - \sum_{j=1}^m u_j + u_i) \leq I(\bar{x}_i)$. The other inequality can be proven similarly.

We know that $l_i \leq I(\bar{x}_i)$ so if $\max(l_i, 1 - \sum_{j=1}^m u_j + u_i) = l_i$, then the inequality holds.

Assume now that $\max(l_i, 1 - \sum_{j=1}^m u_j + u_i) = 1 - \sum_{j=1}^m u_j + u_i$. Then, as $l_i \geq 0$, $1 - \sum_{j=1}^m u_j + u_i \geq 0$ and therefore $\sum_{j=1}^m u_j - u_i \leq 1$.

Assume that the inequality does not hold, i.e., $I(\bar{x}_i) < 1 - \sum_{j=1}^m +u_i$. We know that for all $1 \leq j \leq m$ $I(\bar{x}_j) \leq u_i$. Therefore $\sum_{j=1}^m I(\bar{x}_j) = \sum_{j=1, j\neq i}^m I(\bar{x}_j) + I(\bar{x}_i) \leq \sum_{j=1}^m u_j - u_i + I(\bar{x}_i) < \sum_{j=1}^m u_j - u_i + 1 - \sum_{j=1}^m u_j + u_i = 1$.

But as I is a p-interpretation, $\sum_{j=1}^m I(\bar{x}_j)$ must be equal to 1. The contradiction implies $I(\bar{x}_i) \geq 1 - \sum_{j=1}^m u_j + u_i$.

- P' is tight.

We show that for all $1 \leq i \leq m$, every point $a \in P'(\bar{x}_i)$ is *reachable*. By Proposition 1, it is sufficient to prove that the end points of the $P'(\bar{x}_i)$ interval are reachable.

$P'(\bar{x}_i) = [\max(l_i, 1 - \sum_{j=1}^m u_j + u_i), \min(u_i, 1 - \sum_{j=1}^m l_j + l_j)]$. We show that $\max(l_i, 1 - \sum_{j=1}^m u_j + u_i)$ is reachable. Similar reasoning can be applied to show the reachability of the upper bound.

We show that there exists a p-interpretation I such that $I \models P$ and $I(\bar{x}_i) = \max(l_i, 1 - \sum_{j=1}^m u_j + u_i)$. As we have shown that $P \equiv P'$, $I \models P'$.

Two cases need to be considered. First, let $\max(l_i, 1 - \sum_{j=1}^m u_j + u_i) = l_i$. Then $1 - l_i \leq \sum_{j=1}^m u_j - u_i$ As $\sum_{j=1}^m l_j \leq 1$, we get $\sum_{j=1}^m l_j - l_i \leq 1 - l_i \leq \sum_{j=1}^m u_j - u_i$.

But then (by reasoning similar to that in the proof of Theorem 1), there exist numbers $a_1, \ldots, a_m$, such that $a_i = l_i$ and $(\forall 1 \leq j \leq m)(l_j \leq a_i \leq u_j)$ and $a_1 + \ldots + a_m = 1$. But then, let I be a p-interpretation such that $I(\bar{x}_j) = a_j$ for all $1 \leq j \leq m$. Then $I(\bar{x}_i) = l_i$, $\sum_{j=1}^m I(\bar{x}_j) = 1$ and $I \models P$. Therefore $I \models P'$ and $l_i = \max(l_i, 1 - \sum_{j=1}^m u_j + u_i)$ is reachable.

Let now $\max(l_i, 1 - \sum_{j=1}^m u_j + u_i) = 1 - \sum_{j=1}^m u_j + u_i$.

Consider the function $I : dom(V) \to [0, 1]$, such that $I(\bar{x}_i) = 1 - \sum_{j=1}^m u_j + u_i$ and $I(\bar{x}_j) = u_j$ for all $1 \leq j \leq m, j \neq i$. If I is a p-interpretation then $I \models P$ as $l_i \leq 1 - \sum_{j=1}^m u_j + u_i \leq u_i$ and $u_j \in [l_i, u_j]$. To prove that I is a p-interpretation we must show that $\sum_{j=1}^m I(\bar{x}_j) = 1$.

Indeed, $\sum_{j=1}^m I(\bar{x}_j) = \sum_{j=1, j\neq i}^m I(\bar{x}_j) + I(\bar{x}_i) = \sum_{j=1}^m u_j - u_i + 1 - \sum_{j=1}^m u_j + u_i = 1$. This proves the reachability of $1 - \sum_{j=1}^m u_j + u_i = \max(l_i, 1 - \sum_{j=1}^m u_j + u_i)$, which, in turn proves the theorem. □

3 Conditionalization of Interval Probability Distributions

Interval probabilities arise in probabilistic inference (Bayes nets, for instance) or planning (Markov decision processes). Such stochastic models are often built from probabilistic data in the form of joint probability distributions. In order to compute conditional probability tables, we must define conditionalization. Such conditionalization also occurs in the query algebras of probabilistic databases [4, 2, 1]. These applications require efficient computation methods such as those described in this section.

To illustrate what is involved in the conditionalization of ipdfs, consider the following example.

Example 6 *Consider the joint probability distribution P of two random variables v and v' shown in Figure 2. In this example, our goal is to find the ipdf $P' : dom(v) \to$ C[0,1] that best describes the probability distribution of v given that $v' = a$.*

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

P$' = \mu_{v'=a}$(P):

v	l	u
a	0.5	0.643
b	0.357	0.5

Figure 2: Conditionalization of Interval PDFs

As we recall, P is associated with a set of p-interpretations $\{I|I \models P\}$, each p-interpretation representing a possible point *probability distribution. For each p-interpretation $I : dom(v) \times dom(v') \rightarrow [0,1]$ satisfying P, we can find such a p-interpretation $I' : dom(v) \rightarrow [0,1]$, such that $I'(x) = p(x|v' = a)$ as follows: $I'(x) = \frac{I(x,a)}{I(a,a)+I(b,a)}$ ($x \in \{a,b\}$). The ipdf P' we are trying to determine must then be associated with the set of all such p-interpretations I': $\{I'_I : dom(v) \rightarrow [0,1] | I \models P, I'(x) = \frac{I(x,a)}{I(a,a)+I(b,a)}\}$.*

We can therefore describe P' as $P'(x) = [\min_{I \models P} I'_I(x), \max_{I \models P} I'_I(x)]$.

This reasoning leads to the following definition of the *conditionalization* operation.

In order to simplify the definitions below, we employ the following notation. Let $V = (v_1, \ldots, v_n)$ be a sequence of random variables and let $v \in V$ and $V' = V - \{v\}$. Without loss of generality, we further assume that $v = v_n$ and $V' = \{v_1, \ldots v_{n-1}\}$. Now, let $I : dom(V) \rightarrow [0,1]$ be a p-interpretation. Let $X = \{x_1, \ldots x_k\} \subset dom(v)$ and $\bar{y} \in dom(V')$. We define: $I[X](\bar{y}) = \sum_{i=1}^{m} I(\bar{y}, x_i)$. With this notation in mind, we define conditionalization as follows.

Definition 9 *Let $v \in \mathcal{V}$. A* conditionalization constraint*, c, is an expression of the form "$v = \{x_1, \ldots, x_k\}$" where $x_1, \ldots x_k \in dom(v)$. We slightly abuse notation and write $v = x$ instead of $v = \{x\}$.*

Intuitively, conditionalization of a joint probability distribution P over the set of random variables $V = \{v_1 \ldots v_n\}$ under the constraint $v_n = \{x_1, \ldots, x_k\}$ means finding the joint probability distribution of random variables $\{v_1 \ldots v_{n-1}\}$ under the assumption that v_n takes one of the values $\{x_1, \ldots x_k\}$.

Definition 10 *Let $V = (v_1, \ldots, v_n)$ be a sequence of random variables, $v = v_n$ and $V' = V - \{v\}$. Let c be a conditionalization constraint $v = \{x_1, \ldots, x_k\}$.*

The conditionalization operator μ^V takes as input a (consistent) tight interval pdf P over V and conditionalization constraint c and returns an interval pdf $P' = \mu^V(P, c) : dom(V') \rightarrow$ C[0,1] such that[3]

$$P'(\bar{y}) = \left[\min_{I \models P} \left(\frac{I_X(\bar{y})}{\sum_{y' \in dom(V')} I_X(\bar{y'})} \right), \max_{I \models P} \left(\frac{I_X(\bar{y})}{\sum_{y' \in dom(V')} I_X(\bar{y'})} \right) \right].$$

When V and c are fixed, we write $\mu^V(P, c)$ as $\mu_c(P)$.

Example 6 and Definition 10 specify *what* the result of conditionalization operation should represent, however, they do not specify *how to compute* this result.

[3]The denominator in the expression is 0 only if the numerator is also 0. In such cases we default to a value of 0.

Example 7 *Let us continue where Example 6 left off. In order to compute the result of conditionalization* $P' = \mu_{v'=a}(P)$ *we must find the minimum and maximum values of the expressions of the form* $\frac{I(x,a)}{I(a,a)+I(b,a)}$ *for* $x \in \{a, b\}$ *over all p-interpretations* $I \models P$.

Let us determine these bounds for $x = a$. *As the function* $f(y) = \frac{y}{y+c}$ *for* $c \geq 0$ *is monotonically increasing when* $y \geq 0$, *both minimum and maximum values of* $\frac{I(a,a)}{I(a,a)+I(b,a)}$ *correspond to* $I(a, a)$ *assuming its minimum and maximum value in conjunction with* $I(b, a)$ *being the largest (for minimum) and smallest (for maximum values).*

$P(a, a) = [0.3, 0.45]$ *and* $P(b, a) = [0.25, 0.3]$ *specify the upper and lower bounds on* $I(a, a)$ *and* $I(b, a)$ *respectively. Since* P *is tight, all the bounds are reachable. Thus, we could expect that*
$\max \frac{I(a,a)}{I(a,a)+I(b,a)} = \frac{\max(I(a,a))}{\max(I(a,a))+\min(I(a,b))} = \frac{0.45}{0.45+0.25} = \frac{0.45}{0.7} \cong 0.643$ *and*
$\min \frac{I(a,a)}{I(a,a)+I(b,a)} = \frac{\min(I(a,a))}{\min(I(a,a))+\max(I(a,b))} = \frac{0.3}{0.3+0.3} = \frac{0.3}{0.6} = 0.5.$

There is, however, one caveat. While all four upper and lower bounds are reachable, they are reachable independently, *i.e., we do not know up front whether the upper bound for* $P(a, a)$ *is reachable (in a single p-interpretation) togehter with the lower bound of* $P(b, a)$ *and vice versa. This needs to be checked. In our example, it turns out to be the case: p-interpretation* I_1*:* $I_1(a, a) = 0.45$, $I_1(a, b) = 0.2$, $I_1(b, a) = 0.25$, $I_1(b, b) = 0.1$ *satisfies* P *and reaches both* $\max(I(a, a))$ *and* $\min(I(b, a))$ *and p-interpretation* I_2 : $I_2(a, a) = 0.3$, $I_2(a, b) = 0.2$, $I_2(b, a) = 0.3$, $I_2(b, b) = 0.2$ *satisfies* P *and reaches both* $\min(I(a, a))$ *and* $\max(I(b, a))$.

This allows us to conclude that $P'(a) = [0.5, 0.643]$. *Similar reasoning leads to establishing that* $P'(b) = [1 - 0.643, 1 - 0.5] = [0.357, 0.5]$.

Observe that, by Definition 10, in order to compute the result of conditionalization of an ipdf, a number of non-linear optimization problems must be solved. In the example above, the computation of the conditionalization result was simple because the pairs $\min(I(a, a))$, $\max(I(b, a))$ and $\max(I(a, a))$, $\min(I(b, a))$ were simultaneously reachable. In practice, this is not always the case. When simultaneous reachability is unachievable, the optimization problems become more complex. Luckily, even in such cases, the optimization problems have closed form solutions, as described in the following theorem.

Theorem 3 *Let* $V = (v_1, \dots, v_n)$ *be a sequence of random variables,* $v = v_n$ *and* $V' = V - \{v\}$. *Let c be a conditionalization constraint.*

For $\bar{y} \in dom(V')$ *let*
$l[X]_{\bar{y}} = \max\left(\sum_{x \in X} l_{(\bar{y},x)} \ ; \ 1 - \sum_{\bar{y'} \neq \bar{y} \textbf{ or } x' \notin X} u_{(\bar{y'},x')}\right)$
$u[X]_{\bar{y}} = \min\left(1 - \sum_{\bar{y'} \neq \bar{y} \textbf{ or } x' \notin X} l_{(\bar{y'},x')} \ ; \ \sum_{x \in X} u_{(\bar{y},x)}\right)$
Then

$$P'(\bar{y}) = \left[\frac{l[X]_{\bar{y}}}{\min\left(1 - \sum_{x' \notin X} l_{(\bar{y'},x')} \ , \ \sum_{\bar{y^*} \neq \bar{y} \ , \ x \in X} u_{(\bar{y^*},x)} + l[X]_{\bar{y}}\right)}\right. ,$$
$$\left.\frac{u[X]_{\bar{y}}}{\max\left(\sum_{\bar{y^*} \neq \bar{y} \ , \ x \in X} l_{(\bar{y^*},x)} + u[X]_{\bar{y}} \ , \ 1 - \sum_{x' \notin X} u_{(\bar{y'},x')}\right)}\right]$$

Proof (*sketch*). We sketch here the proof of the theorem for the case when the conditionalization constraint c is of the form $v = x$.

Let $\bar{y} \in dom(V')$. We need to find $P'(\bar{y}) = [l'_{\bar{y}}, u'_{\bar{y}}]$
Consider the problem of finding the lower bound $l'_{\bar{y}}$ of P'.
As follows from Definition 10 of conditionalization,

$$l_{\bar{y}} = \min_{I \models P} \left(\frac{I(\bar{y}, x)}{\sum_{\bar{y}' \in dom(V')} I(\bar{y}', x)} \right) = \min_{I \models P} \left(\frac{I(\bar{y}, x)}{I(\bar{y}, x) + \sum_{\bar{y}' \in dom(V'), \bar{y}' \neq \bar{y}} I(\bar{y}', x)} \right).$$

This minimization problem can be simplified to $\min(f(w,z)) = \min(\frac{w}{w+z})$. Given the additional constraints on the relationship of w and z imposed by p-interpretation origins of these variables, the minimum of this function can be computed by first computing the minimum w' for w and then finding the maximum z satisfying the ensuing constraints when $w = w'$. Substituting $w = I(\bar{y}, x)$ and $z = \sum_{\bar{y}' \in dom(V'), \bar{y}' \neq \bar{y}} I(\bar{y}', x)$ we get

$$\min_{I \models P} \left(\frac{I(\bar{y}, x)}{I(\bar{y}, x) + \sum_{\bar{y}' \in dom(V'), \bar{y}' \neq \bar{y}} I(\bar{y}', x)} \right) = \frac{\min_{I \models P}(I(\bar{y}, x))}{\min_{I \models P}(I(\bar{y}, x)) + \max_{I \models P}(\sum_{\bar{y}' \in dom(V'), \bar{y}' \neq \bar{y}} I(\bar{y}', x))}.$$

Now, we find $\min_{I \models P}(I(\bar{y}, x))$ and $\max_{I \models P}(\sum_{\bar{y}' \in dom(V'), \bar{y}' \neq \bar{y}} I(\bar{y}', x))$.

From the proof of Theorem 2 we know that $\min_{I \models P}(I(\bar{x})) = \max(l_{\bar{x}}, 1 - \sum_{\bar{x}' \neq x} u_{\bar{x}'})$, i.e., $\min_{I \models P}(I(\bar{y}, x) = \max(l_{(\bar{y},x)}, 1 - \sum_{\bar{x}' \neq (\bar{y},x)} u_{\bar{x}'} = l[\{x\}]_{\bar{y}})$.

Now, fixing the value of $I(\bar{y}, x)$ to $l[\{x\}]_{\bar{y}}$, the maximum of the sums of the $I(\bar{y}, x)$ of the remaining $(\bar{y}', x)$ $(\bar{y} \neq \bar{y}')$ rows of $dom(V)$ will be the minimum of the following two quantities (the other being unreachable):

- the sum of their respective upper bounds according to P: $\sum_{\bar{y'} \in dom(V'), \bar{y'} \neq \bar{y}} u_{(\bar{y'}, x)}$;
- 1 minus the sum of lower bounds of all other rows of $dom(V)$ according to P:
  $1 - \sum_{\bar{y}' \in dom(V')} \sum_{x' \neq x} l_{(\bar{y}', x')} - l[\{x\}]_{\bar{y'}}$.

These formulas, when combined, give us the desired result. Similar reasoning establishes the formula for the upper bound. □

4 Related Work and Conclusions

Imprecise probabilisties have attracted the attention of researchers for quite a while now, as documented by the Imprecise Probability Project [8]. The seminal work of Walley [10] made the case for interval probabilities as the means of representing uncertainty. In this book, Walley talks about the computation of conditional probabilities of events. As discussed in Section 1, his semantics is quite different from ours.

On the other hand, Weichselberger [11] extends the Kolmogorov axioms of probability theory to the case of interval probabilities. As it builds on Kolmogorov probability theory, the interval probability semantics is defined for a σ-algebra of random events. Weichselberger completes his theory with the definition of conditional probability. Our semantics can be viewed as extending the semantics of Weichselberger [11] to the case of probability distributions of *discrete random variables*. Our notion of consistent interval pdfs corresponds to Weichselberger's *R-probabilities* and our tight interval pdfs correspond to his *F-probabilities*. However, his definition of conditional probability applies to a different structure than ours. Note that this is also true about the work on conditional probabilities by Walley [10]. Dekhtyar, Ross and Subrahmanian in [3] developed a specialized semantics for probability distributions used in their Temporal Probabilistic Database model. In particular, they defined the notions of consistency and tightness of interval probability distributions. Our semantics generalizes theirs. One other instance of the possible world semantics for interval probabilities occurs in Givan, Leach and Dean's discussion of Bounded Paramenter Markov Decision Processes [6].

Conditionalization as an operation in a probabilistic database model had first been considered by Dey and Sarkar [4]. Dekhtyar, Goldsmith and Hawkes also use this operation in their Semistructured Probabilistic Algebra [2]. In both works, conditionalization is performed on *point probability distributions* of discrete random variables, and the operation itself is fairly striaghtforward. Interval probabilities have attracted the attention of a number of researchers in databases [9, 3, 5], but the database models proposed did not include the conditionalization operation. The work described here is part of the research that lead to the extension of the Semistructured Probabilistic Database model of [2] to the case of interval probabilities [1].

Acknowledgements

We would like to thank Greg Wasilkowski and Yuri Prostota for productive discussions, David Harmanec for pointing us toward Walley's work, and Andy Klapper for helping us parse some of the definitions therein. All misunderstandings, however, are the sole responsibility of the authors.

References

[1] A. Dekhtyar and J. Goldsmith. (2002). Semistructured Models for Interval Probabilities. *UK CS Tech Report, submitted to conference*.

[2] A. Dekhtyar, J. Goldsmith, S.R. Hawkes. (2001) Semistructured Probabilistic Databases, in *Proc. SSDBM'2001*, pp. 36–45.

[3] A. Dekhtyar, R. Ross, V.S. Subrahmanian. (2001) Temporal Probabilistic Databases, I: Algebra, *ACM Transactions on Database Systems*, Vol. 26, 1, pp. 41–95.

[4] D. Dey, S. Sarkar. (1996) A Probabilistic Relational Model and Algebra, *ACM Transactions on Database Systems*, Vol. 21, 3, pp. 339–369.

[5] T. Eiter, T. Lukasiewicz, M. Walter. (2001) A data model and algebra for probabilistic complex values *Annals of Mathematics and Artificial Intelligence*, Vol. 33 No. 2-4, pp. 205–252.

[6] R. Givan, S. Leach, T. Dean. (2000) Bounded-Parameter Markov Decision Processes, *Artificial Intelligence*, Vol. 122, 1-2, pp. 71–109.

[7] Kleiter, G. (1996). Propagating imprecise probabilities in Bayesian networks. *Articial Intelligence* vol. 88, pps. 143–161.

[8] H. E. Kyburg, Jr. (1998). Interval-valued probabilities. In G. de Cooman, P. Walley, and F. G. Cozman, editors, *Imprecise Probabilities Project*. 1998. Available from http://ippserv.rug.ac.be/.

[9] V.S. Lakshmanan, N. Leone, R. Ross and V.S. Subrahmanian. (1997) ProbView: A Flexible Probabilistic Database System. *ACM Transactions on Database Systems*, Vol. 22, Nr. 3, pps 419–469, Sep. 1997.

[10] Walley, P. (1991). Statistical Reasoning with Imprecise Probabilities. Chapman and Hall, 1991.

[11] Weichselberger, K. (1999). The theory of interval-probability as a unifying concept for uncertainty. *Proc. 1st International Symp. on Imprecise Probabilities and Their Applications*, July 1999.