数理解析研究所講究録 1420 巻 2005 年 163-173

A Classification of the Probabilistic Reasoning given Distribution Evidence and Kullback-Leibler Information

Toshiyasu Matsushima 松嶋敏泰
Waseda University 早稲田大学

November 18, 2004

Abstract

The probabilistic reasoning given distribution evidence, virtual evidence, indirect evidence or likelihood have been investigated in previous research. In this paper, we classify the reasoning into two types and define each type of the reasoning by mathematical formulas. From the definition, we show that the first type of reasoning is solved by the minimization of Kullback-Leibler(K-L) information under marginal constraints and the second type is calculated by the ordinary probabilistic reasoning methods such as Belief Propagation(BP). We also show that the Iterative Scaling Procedure(ISP) is applied to the first type reasoning. Moreover, we propose an efficient propagation algorithm, which are based on ISP, for the reasoning on Junction trees. Both the space and the time complexities of the proposed algorithm are lower than that of the previous research.

1 Introduction

A probabilistic reasoning problem is defined by its input and the target output. The input of probabilistic reasoning[Jensen 1996] [Pearl 1988] is a joint distribution $P(X_1,\dots,X_n)$ [1] and information about deterministic values of some random variables $X_j = x_j \ j \in I_C = \{1,2,\dots,k\} \subset I = \{1,2,\dots,n\}$ called evidence. The target output is a joint distribution $P_{out}(X_1,\dots,X_n) = P(X^{I-I_C}, X^{I_C} = x^{I_C})$ where $X^{I_C} = x^{I_C}$ denotes $(X_1 = x_1,\dots,X_k = x_k)$ or the marginal distributions of P_{out}. Thus the probabilistic reasoning problem is defined as a deduction of the posterior probability $P(X^{I-I_C}|X^{I_C} = x^{I_C})$ or $P(X^{I-I_C}, X^{I_C} = x^{I_C})$ from a prior probability and the information about the occurrence of some random variables $X_j = x_j \ j \in I_C$. The correctness of typical reasoning algorithms such as Belief Propagation (BP) or HUGIN has been proved by how the algorithms can calculate the target posterior distribution correctly.

The input information in ordinary probabilistic reasoning is given by deterministic values of some random variables $X_j = x_j \ j \in I_C$, while some previous research has investigated reasoning in which the input information is given by distributions of some random variables $P^*(X_j) \ j \in I_C$. This kind of evidence is called soft evidence, distribution evidence, virtual evidence, indirect evidence or likelihood in the previous research [Pearl 1990][Valtorta *et al.* 2002].

There was no clear classification between these evidences. We think the reasoning using such kind of evidence can be classified into two types from the viewpoint of the desired character of the target output distributions. In the first type of reasoning, which we call Type 1 reasoning, the given distributions $P^*(X_j) \ j \in I_C$ coincide with the marginal of the output distribution $P_{out}(X_j) = \sum_{i \neq j} P_{out}(X_1,\dots,X_n)$. In the second type, which we call Type 2 reasoning, the given distribution does not always coincide with the marginal of the output distribution.

Many kinds of reasoning methods have been proposed for Type 1 reasoning. A typical method for Type 1 reasoning is based on Jeffry's Rule[Pearl 1990]. Another typical reasoning method is based on the principle of least change or the minimum divergence principle[Wen 1990] that minimizes the divergence

[1] In probabilistic reasoning, the joint distribution is represented by some graph. This means that the joint distribution factors into a product of several functions of some subset of random variables. We investigate this in Section 4.

between an input distribution P_{in} and the output distribution P_{out} under some condition. On the other hand, the ordinary probabilistic reasoning algorithms such as the BP or the HUGIN propagation have been used for Type 2 reasoning.

The previous research only proposed reasoning methods individually, but we have not seen any research discussing the relationship between them or proving the mathematical justification of them. The justification of the methods has been only given by qualitative and intuitive explanation in the previous research.

We think the lack of mathematical justification is caused by a lack of a mathematical definition of the reasoning problems. We have not seen any research defining Type 1 reasoning problems by mathematical formulas. If there is no mathematical definition of the input information and the target output, we cannot evaluate the justification of reasoning algorithms.

The first objective of this paper is to give definitions of Type 1 and Type 2 reasoning problems by mathematical formulas. We clarify the difference between Type 1 and Type 2 reasoning by these definitions. The defined reasoning problems include ordinary probabilistic reasoning problems.

Secondly we show the correct reasoning under the definition. Type 1 reasoning is solved by the minimization of Kullback-Leibler(K-L) information under the marginal distribution constraints represented by Type 1 evidence. Type 2 reasoning is exactly calculated by ordinary probabilistic reasoning algorithms such as BP. Since we can use ordinary reasoning algorithms for Type 2 reasoning, we mainly investigate Type 1 reasoning in the following sections.

Thirdly we propose basic procedures for Type 1 probabilistic reasoning. An efficient procedure called the Iterative Scaling Procedure(ISP) [Csiszar 1975][Ireland and Kullback 1968] can be applied to the reasoning procedure. Although some previous research applied ISP to the method of reasoning given distribution evidence, we deduce the procedure using ISP from a simple assumption and the definition without intuitive concepts.

Finally an efficient algorithm of Type 1 reasoning on Junction Trees(JT) [Aji and Mcliece 2000] [Jensen 1996] is proposed. The big clique algorithm using ISP was proposed in the previous research [Valtorta *et al.* 2002]. An effective implementation of ISP for the maximum likelihood estimation on contingency tables was also investigated in the previous research [Jirousek and Preucil 1995]. The complexities of the previous algorithms are higher than that of the proposed algorithm.

2 Formalization of Type 1 and Type 2 probabilistic reasoning

First, we define Type 1 and Type 2 probabilistic reasoning. Let X_i $i \in I$ and E_j $j \in I_C \subset I$ be discrete random variables for sake of brevity and E_j is called implicit evidence.

We assume each piece of implicit evidence E_j gives us the information only about X_j not about the other X_i $i \in I - \{j\}$ [2]. Formally, the above mentioned condition is represented by the following assumption.

Assumption 1 [3] *Each piece of implicit evidence E_j and every X_i $i \in I - \{j\}$ are conditionally independent given X_j as follows:*

$$P(X_1, \dots, X_n, E^{I_C}) = \frac{P(X_1, \dots, X_n) \prod_{j \in I_C} P(X_j, E_j)}{\prod_{j \in I_C} P(X_j)}, \tag{1}$$

where E^{I_C} denotes $(E_1, \dots, E_k)$.

Now, we define Type 1 probabilistic reasoning.

Definition 1 *Type 1 probabilistic reasoning is defined by the following input and output. The input is given by a distribution $P(X_1, \dots, X_n)$ and the information of X_j $j \in I_C$ as $P^*(X_j) = \sum_{i \neq j} P(X_1, \dots, X_n|$*

[2] It is easy to extend this assumption to the assumption that each piece of implicit evidence gives us the information about $(X_{j_1}, \dots, X_{j_l})$.

[3] This assumption is identical with the assumption of the previous research[Pearl 1990].

$E^{I_C} = e^{I_C}) = P(X_j|E^{I_C} = e^{I_C})$, which is the marginal distribution of X_j given evidence $E^{I_C} = e^{I_C}$, where $E^{I_C} = e^{I_C}$ denotes $(E_1 = e_1, \ldots, E_k = e_k)$. The target output $P_{out}$ is the distribution $P(X_1, \ldots, X_n|E^{I_C} = e^{I_C})$ or the marginal distributions of $P_{out}$.

The output of the defined Type 1 probabilistic reasoning $P(X^I|E^{I_C} = e^{I_C})$ differs from the output of the ordinary probabilistic reasoning, which is the conditional distribution $P(X^{I-I_C}|X^{I_C} = x^{I_C})$. However the conditional distribution $P(X^I|E^{I_C} = e^{I_C})$ includes $P(X^{I-I_C}, X^{I_C} = x^{I_C})$ as a special case. If the distribution $P(X_j|E^{I_C} = e^{I_C})$ is the point mass in $X_i = x_i$, i.e., $P(X_j = x_i|E^{I_C} = e^{I_C}) = 1$, the information from the implicit evidence e_j is the same as "x_i occurred", i.e., $X_i = x_i$. In this case, the defined output distribution is identical with $P(X^{I-I_C}, X^{I_C} = x^{I_C})$. Thus, the defined Type 1 probabilistic reasoning includes the ordinary probabilistic reasoning as a special case.

From Definition 1, the given distributions $P^*(X_j)$ $j \in I_C$ coincide with the marginal of the output distribution $P_{out}(X_j) = \sum_{i \neq j} P(X_1, \ldots, X_n|E^{I_C} = e^{I_C})$. This satisfies the requirement of the previous research on Type 1 reasoning.

Next, we define Type 2 probabilistic reasoning.

Definition 2 *Type 2 probabilistic reasoning is defined by the following input and output. The input is given by a distribution $P(X_1, \ldots, X_n)$ and information $P^{**}(X_j) = \alpha P(X_j, E_j = e_j)$ or $P(X_j, E_j = e_j)/P(X_j)$ in Formula(1). The target output $P_{out}$ is the distribution $P(X_1, \ldots, X_n|E^{I_C} = e^{I_C})$ or the marginal distributions of $P_{out}$.*

The difference between two types of reasoning problems is the information of X_j $j \in I_C$ given by implicit evidence. Thus, $P^{**}(X_j)$ does not always coincide with the marginal of the output distribution $P_{out}(X_j) = \sum_{i \neq j} P(X_1, \ldots, X_n|E^{I_C} = e^{I_C})$. We can easily prove that ordinary probabilistic reasoning algorithms deduce the output distribution of Type 2 reasoning. That is the reason why the HUGIN algorithm can be applied to the reasoning for indirect evidence or likelihood, i.e., Type 2 evidence.

Remark 1 *In this problem setting, it is important that we know the marginal distributions $P(X_1, \ldots, X_n)$ as the input distribution but do not need the whole joint distribution $P(X_1, \ldots, X_n, E^{I_C})$. Under Assumption 1, we can determine the target output distribution $P(X_1, \ldots, X_n|E^{I_C} = e^{I_C})$ from the information $P(X_j|E_j = e_j)$ and $P(X_1, \ldots, X_n)$ without $P(X_1, \ldots, X_n, E^{I_C})$. Although the problem is defined on the probability space of $(X_1, \ldots X_n, E^{I_C})$, we can treat the problem as being only on the space of $(X_1, \ldots, X_n)$, which is the same space as for ordinary probabilistic reasoning.*

The output distributions calculated by Type 1 probabilistic reasoning are interpreted as generalized posterior distributions given marginal distributions $P^*(X_j)$ instead of given strict values $X^{I_C} = x^{I_C}$. Generalized posterior distributions play the same role as posterior distributions do in statistical inference.

3 A Basic procedure for Type 1 probabilistic reasoning

3.1 Relationship between the output distribution and a prior distribution

We investigate the property of the output distribution deduced by the defined Type 1 probabilistic reasoning. The relationship between the output distribution and a prior distribution is shown by the following lemma and theory.

Lemma 1 *Under Assumption 1, the output distribution $P_{out} = P(X_1, \ldots, X_n|E^{I_C} = e^{I_C})$ that is deduced from an input distribution $P(X_1, \ldots, X_n)$ and information $P^*(X_j) = P(X_j|E^{I_C} = e^{I_C})$ $j \in I_C$ by Type 1 reasoning is given by*

$$P_{out} = \alpha P(X_1, \ldots, X_n) \prod_{j \in I_C} \beta(X_j), \tag{2}$$

where $\beta(X_j) > 0$.

Proof:

$$\begin{aligned} P_{out} &= P(X_1,\dots,X_n|E^{I_C}=e^{I_C}) && (3)\\ &= \alpha P(X_1,\dots,X_n,E^{I_C}=e^{I_C}) && (4)\\ &= \alpha\frac{P(X_1,\dots,X_n)\prod_{j\in I_C}P(X_j,E_j=e_j)}{\prod_{j\in I_C}P(X_j)} && (5)\\ &= \alpha P(X_1,\dots,X_n)\prod_{j\in I_C}\beta(X_j). && (6)\end{aligned}$$

Formula(5) is given by Assumption 1.

Remark 2 *If $P(y)\neq 0$ then the conditional probability $P(x|y)=P(x,y)/P(y)$ can be defined. So the region $R(x^{I_C})$ of deterministic value of X_j for the evidence of ordinary probabilistic reasoning is restricted as follows:*

$$R(x^{I_C})=\{x^{I_C}|P(x^{I_C})\neq 0\}. \tag{7}$$

In a similar fashion, the region of the value of the probability given as Type 1 evidence is restricted as follows:

$$\begin{aligned} &R(P^{I_C})=\\ &\quad\{P^{I_C}\mid \exists\beta(X_1)>0\cdots\exists\beta(X_k)>0\ \forall l\in I_C\\ &\quad P(X_l)=\sum_{X\neq X_l}\alpha P(X_1,\dots,X_n)\prod_{j\in I_C}\beta(X_j)\},\end{aligned} \tag{8}$$

where $P^{I_C}=(P(X_1),\dots,P(X_k))$.

If $P^{I_C}\in R(P^{I_C})$ then the generalized posterior distribution or the generalized conditional probability given P^{I_C} can be defined. It is regarded as a generalization of the condition under which the ordinary conditional distribution can be defined.

An important characteristic of the output distribution is shown by the following theorem.

Theorem 1 *Let M_C be the set of the distributions on the random variables $X_1,\dots,X_n$ that satisfy the marginal condition $P(X_j)=P^*(X_j)$ $j\in I_C$ and $P^{I_C}\in R(P^{I_C})$. Under Assumption 1, the output distribution $P_{out}=P(X_1,\dots,X_n|E^{I_C}=e^{I_C})$ that is deduced by Type 1 reasoning is given by*

$$P_{out}=\arg\min_{P\in M_C} I(P\|P_{in}), \tag{9}$$

where P_{in} is a prior distribution $P(X_1,\dots,X_n)$ and $I(\cdot\|\cdot)$ is Kullback-Leibler(K-L) information.

Proof:

Let P_M be the distribution as follows:

$$P_M=\arg\min_{P\in M_C} I(P\|P_{in}). \tag{10}$$

We consider the following $T_{x_i}(X_1,\dots,X_n)$ as $T(x)$ in Theorem 2.1 of the paper[Kullback 1959].

$$T_{x_i}(X_1,\dots,X_n)=\begin{cases}1 & X_i=x_i\\ 0 & X_i\neq x_i\ .\end{cases} \tag{11}$$

From Theorem 2.1 of the paper, P_M is given by

$$P_M=\frac{e^{\sum_{i=1}^{k}\tau_{x_i}}P(x_1,\dots,x_n)}{\sum_{x_1}\cdots\sum_{x_n}e^{\sum_{i=1}^{k}\tau_{x_i}}P(x_1,\dots,x_n)}. \tag{12}$$

Let us set $e^{\tau x_i} = \beta'(x_i)$.

$$P_M(x_1,\ldots,x_n) = \alpha' P(x_1,\ldots,x_n) \prod_{j \in I_C} \beta'(x_j), \tag{13}$$

where

$$P_M(x_j) = P^*(x_j)\ j \in I_C. \tag{14}$$

From Lemma 1 and Definition 1, P_{out} *is represented by*

$$P_{out}(x_1,\ldots,x_n) = \alpha P(x_1,\ldots,x_n) \prod_{j \in I_C} \beta(x_j), \tag{15}$$

where

$$P_{out}(x_j) = P^*(x_j)\ j \in I_C. \tag{16}$$

From the uniqueness of $\beta'(x_j)$ *and* $\beta(x_j)$*, we obtain the following formula and the theorem can be proved.*

$$P_{out}(x_1,\ldots,x_n) = P_M(x_1,\ldots,x_n) \tag{17}$$

The theory shows that the distribution calculated by Type 1 probabilistic reasoning is the distribution that is closest to the prior distribution with K-L information under the restriction of marginal distributions.

Some previous research proposed the reasoning methods based on the principle of least change or the minimum divergence principle[Wen 1990] that minimizes the divergence between prior distribution P_{in} and output distribution P_{out} under some condition. However the correctness of the principle also has not been justified in the research. There are a lot of measures of the divergence between two distributions. For example, if we use K-L information as the divergence, which divergence is correct, $I(P_{in}\|P_{out})$ or $I(P_{out}\|P_{in})$? The selection of the measure has been still supported by a qualitative concept or one's intuition in the research. In this paper, justification of the minimum divergence principle can be proved from the definition of Type 1 probabilistic reasoning and Assumption 1.

On the other hand, there was some previous research investigating the distribution given by minimizing K-L information. The paper[Kullback 1959] used in the proof of Theorem 1 is one example of the previous research. The paper showed that the distribution given by minimizing K-L information under some linear restriction is represented by the product of some parameters and a prior distribution as Formula (2). Inversely, the previous paper[Skyrms1985] claimed that if the target distribution is assumed as the product of some parameters and a prior distribution then the distribution can be deduced by minimizing K-L information under some restriction.

Since Type 1 reasoning was not defined by mathematical formalization in the previous research, the property of the output distribution was unclear. However, from Definition 1 and Lemma 1, we show that the output distribution of Type 1 reasoning can be represented by the product of some parameters and a prior distribution. Thus we can also show that the output distribution can be deduced by minimizing K-L information in Theorem 1.

3.2 A basic procedure for Type 1 probabilistic reasoning and ISP

Type 1 probabilistic reasoning problem shown in Theorem 1 is regarded as one of the conditional optimization problems. The computational complexity for calculating an optimum solution in a conditional optimization problem is generally very high. However, Iterative Proportional Fitting Procedure (IPFP) or the Iterative Scaling Procedure (ISP) can be applied to the procedure of the defined Type 1 probabilistic reasoning. ISP is used for computing the maximum likelihood estimators (MLEs) in a probabilistic model of a contingency table[Ireland and Kullback 1968] under the condition that some marginal sums are given. ISP is also applied to Type 1 probabilistic reasoning.

[Procedure 1: ISP]

```
begin
$P(X_1,\dots,X_n) := P_{in}(X_1,\dots,X_n)$;
$i := 1$;
while $\exists_{j \in I_C} P(X_j) \neq P^*(X_j)$ do
    begin
    $j := i \bmod |I_C|$;
    $P(X_1,\dots,X_n) := P(X_1,\dots,X_n)\frac{P^*(X_j)}{P(X_j)}$;
    $i := i + 1$;
    end
$P_{out}(X_1,\dots,X_n) := P(X_1,\dots,X_n)$;
end
```

Lemma 2 *If $P^{I_C *} \in R(P^{I_C})$ then Procedure 1 halts and the value calculated by Procedure 1 converges to $P(X_1,\dots,X_n|E^{I_C} = e^{I_C})$.*

Proof: *It is obvious from Theorem 1 and the property of ISP[Csiszar 1975][Ireland and Kullback 1968].*

ISP is a very simple iterative procedure for calculating generalized posterior distributions. ISP renews the distribution by adjusting its marginal probability to each restricted marginal value P^* at each cycle. ISP repeats this renewal iteratively until the marginals of the calculated distribution converge to the restricted values.

Procedure 1, i.e., the ISP for Type 1 probabilistic reasoning, differs from the ISP for MLE at several points. Each joint probability $P(x_1,\dots,x_n)$ corresponds to a cell of the contingency table of the ISP for MLE. All cells of the contingency table are set to a constant at the first stage in the ISP for MLE. The given marginal distributions correspond to the marginal sums of given data in the ISP for MLE.

The application of Jeffry's Rule to Type 1 reasoning was proposed by the previous research[Pearl 1990]. Procedure 1 is identical with Jeffry's Rule in the case given one piece of Type 1 evidence. Some previous research[Valtorta *et al.* 2002] applied ISP to the reasoning method. However the research has only proposed the method without defining the target output distribution. Although the previous research has only given qualitative and intuitive explanations for applying ISP to the reasoning method, we deduce procedure 1, which is the procedure using ISP, from Assumption 1, Definition 1 and Theorem 1 without intuitive concepts.

4 An Efficient procedure on Junction Trees

4.1 Probability model of Junction Trees

A Junction Graph/Tree is defined by a clique node set $S^N = \{N_1, N_2, \dots N_{n_N}\}$, an intersection node set $S^D = \{D_1, \dots, D_{n_D}\}$ and the neighboring node set $S^N(D_m)$ of every intersection node $D_m, m = 1,\dots,n_D$.

Each intersection node is connected to all clique nodes in its neighboring node set with arcs in a Junction Graph/Tree. A Junction Tree(JT) is applied to the representation of the probability model whose joint distribution factors into a product of several local functions of some subset of random variables. A typical type of joint distributions represented by JTs are shown as follows:

$$P(X_1,\dots,X_n) = \frac{P(N_1)P(N_2)\cdots P(N_{n_N})}{P(D_1)\cdots P(D_{N_D})}, \tag{18}$$

where $P(N_l) = P(X_{i_1(l)},\dots,X_{i_{n(l)}(l)})$ and $P(D_m) = P(X_{i_1(m)},\dots,X_{i_{n(m)}(m)})$.

$t(N_l) = \{X_{i_1(l)},\dots,X_{i_{n(l)}(l)}\}$ $l \in \{1,\dots,n_N\}$ and $t(D_m) = \{X_{i_1(m)},\dots,X_{i_{n(m)}(m)}\}$ $m \in \{1,\dots,n_D\}$ are called clique elements and intersection elements respectively. Abbreviate $t(N_l)$, $t(D_m)$ to N_l, D_m respectively. The distributions that can be represented by BN are included in this type of distributions.

4.2 A propagation algorithm on Junction Trees

We propose a new propagation algorithm on JTs for calculating the marginal distributions of the output distribution: $P_{out}(N_l) = \sum_{X \notin N_l} P(X_1, \cdots, X_n | E^{I_C} = e^{I_C}), l = 1, \ldots, n_N$. The JT used for the calculation of Type 1 reasoning is the same as that of ordinary probabilistic reasoning. So the JT does not have the cliques including implicit evidence E_j.

Before the propagation algorithm is explained, several terms are defined. First, restricted intersection node(RIN) is defined. If the element of an intersection node is equivalent to a restricted random variable as $D_m = \{X_j\}$ $j \in I_C$, the intersection node is called a restricted intersection node(RIN). If there does not exist an intersection node satisfying $D_m = \{X_j\}$ $j \in I_C$, the RIN corresponding to every such restricted random variable X_j is produced and connected to an arbitrary clique node N_l satisfying $X_j \in N_l$ [4]. The set of all RINs is denoted by S_{RIN}.

The restricted tree(RT) of a JT is defined as the smallest subtree whose leaves are all RINs in the JT [5]. All clique nodes on a RT are numbered in order of the depth first search from an arbitrary root as in Fig 1. The intersection node between a clique node N_u and N_v is denoted by $D_{u,v}$. Each node may have multiple numbers. We numbers the clique nodes and the intersection nodes in a RT with a view to simplifying and clarifying the procedure and the proof of Theorem 2. So the indexes of nodes in a RT are different from those in Formula (18). The maximum number of the numbered cliques in a RT is denoted by l_{RT}.

Example 1 *Let a prior joint distribution be*

$$
\begin{aligned}
P(X_1, \ldots, X_n) = & \\
& \frac{P(X_1, X_4)P(X_4, X_5)P(X_5, X_6, X_7)P(X_2, X_6)}{P(X_2)P(X_3)P(X_4)} \\
& \frac{P(X_2, X_8)P(X_3, X_7)P(X_3, X_9)}{P(X_5)P(X_6)P(X_7)}. \qquad (19)
\end{aligned}
$$

Let the restricted random variables, i.e., the variables whose distributions are given as Type 1 evidence be X_1, X_2, X_3. The RINs in the original JT of the joint distribution mentioned above are $\{X_2\}, \{X_3\}$ and there is no intersection node corresponding to X_1. So we produce the new RIN $\{X_1\}$ and connect it to the clique node $\{X_1, X_4\}$ as in Fig 1.

The restricted tree(RT) of the JT is shown in Fig 1. The leaves of the RT are the intersection nodes $\{X_1\}, \{X_2\}, \{X_3\}$, which are all RINs. The RT is constructed by deleting the clique nodes $\{X_2, X_8\}$, $\{X_3, X_9\}$ from the original JT.

[The strategy of propagation]

1. Repeat the propagation of Procedure 2 on the RT until the values of all cliques converge to some value.
2. Propagate messages from the RT to the whole JT by an ordinary probabilistic reasoning algorithm.

[Procedure 2]

```
begin
$i := \min\{k | D_{k,k+1} \in S_{RIN}\}$;
while $\exists_{D \in S_{RIN}} P(D) \neq P^*(D)$ do
    begin
    $u := i \bmod l_{RT}$;
    $v := i + 1 \bmod l_{RT}$;
    if $D_{u,v} \in S_{RIN}$ then $P(D_{u,v}) := P^*(D_{u,v})$
```

[4] When we connect the produced RINs to clique nodes, from the viewpoint of complexity, we should select the clique nodes as the extended JT includes the smallest restricted tree.

[5] In the case where the given JT is divided into subtrees by disconnecting every middle RIN, we can calculate the target distribution by applying the following propagation algorithm to the RT of each divided sub JT individually.

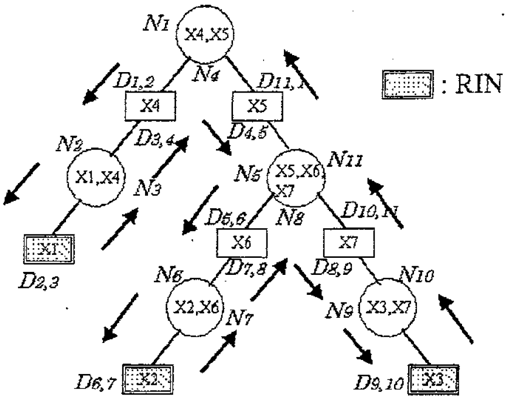

Figure 1: Numbered clique nodes and intersection nodes on a restricted tree and the route of propagation, where the black box nodes are RINs.

```
        else $P(D_{u,v}) := \sum_{X \notin D_{u,v}} P(N_u)$ ;
        $P(N_v) := P(N_v) \frac{P(D_{u,v})}{\sum_{X \notin D_{u,v}} P(N_v)}$ ;
        $i := i + 1$ ;
        end
$P_{out}(N) := P(N)$;
end
```

Theorem 2 *If $P^{I_C*} \in R(P^{I_C})$ then Procedure 2 halts and the value calculated by Procedure 2 converges to $P(N|E^J = e^J)$ in every clique.*

Proof: *See Appendix.*

4.3 Comparisons between Procedure 2 and previous research

The propose procedure propagates messages in numerical order of the clique numbers. The calculation at each clique is simple and similar to the HUGIN propagation. Although the HUGIN propagation stops after a round trip between leaves and the root, i.e., Collect Evidence and Distribution Evidence, Procedure 2 repeats the cycle until the calculated values converge to the target values. Needless to say, the HUGIN algorithm can be applied to only ordinary reasoning and Type 2 reasoning, but not to Type 1 reasoning.

The space complexity of Procedure 2 is $O(\sum_{N \in S^N_{RT}} \prod_{X \in N} |X|)$ where S^N_{RT} is the set of all clique nodes in the RT and $|X|$ denotes the number of values in X. The time complexity of multiplication in one cycle of Procedure 2 is $2(\sum_{N \in S^N_{RT}} \prod_{X \in N} |X| + \sum_{D \in S^D_{RT}} \prod_{X \in D} |X|)$ where S^D_{RT} is the set of all intersection nodes in the RT. The time complexity of addition in one cycle of Procedure 2 is $2(\sum_{N \in S^N_{RT}} \prod_{X \in N} |X|)$.

The space complexity for variables of Procedure 1 is $O(\prod_{i \in I} |X_i|)$. The time complexity of multiplication in one cycle [6] of Procedure 1 is $|I_C| \prod_{i \in I} |X_i|$. So both the space and the time complexities of Procedure 2 are extremely lower than those of Procedure 1, i.e., ISP.

The big clique algorithm, which applies ISP to the calculation of soft evidence reasoning on JTs, was proposed in the previous research[Valtorta *et al.* 2002]. The algorithm uses the big clique that includes the whole RT for the calculation. The algorithm applies an ISP directory to the big clique. So the space complexity of the algorithm is more than $O(\prod_{X \in \bigcup_{N \in S^N_{RT}} N} |X|)$. The time complexity of multiplication

[6] One cycle corresponds to $|I_C|$ loops from $j = 1$ to $|I_C|$ in Procedure 1.

in one cycle of the algorithm is $|I_C| \prod_{X \in \bigcup_{N \in S^N_{RT}} N} |X|$ and the time complexity of addition in one cycle is $|I_C| \prod_{X \in \bigcup_{N \in S^N_{RT}} N} |X|$. Thus both the space and the time complexities of the previous algorithm are higher than those of Procedure 2.

An effective algorithm, which applies ISP for calculating maximum likelihood on a contingency table, was proposed in a previous paper[Jirousek and Preucil 1995]. The message of each RIN is propagated on a graph in order of satisfying the running intersection property. This method is interpreted as the following procedure. Let a RIN be a root of the JT representing a prior probabilistic model. The algorithm propagates the marginal restriction of the root RIN to the whole JT in the same way as the calculation of a clique in the HUGIN propagation. Next, let another RIN be a root of the JT. And the algorithm repeats the same manner of calculation until the values converge.

So the time complexity of multiplication in one cycle of the algorithm is $|I_C|(\sum_{N \in S^N} \prod_{X \in N} |X| + \sum_{D \in S^D} \prod_{X \in D} |X|)$. The time complexity of addition in one cycle of the algorithm is $|I_C|(\sum_{N \in S^N} \prod_{X \in N} |X|)$. Even if the RT of a JT is the same as the JT, the time complexity of the previous algorithm is higher than that of Procedure 2. The space complexity of the algorithm is $O(\sum_{N \in S^N_{RT}} \prod_{X \in N} |X|)$. Thus the space complexity of the algorithm is the same as that of the proposed algorithm.

5 CONCLUSION

We defined Type 1 and Type 2 probabilistic reasoning problem by mathematical formulas. The correct reasoning under the definition of Type 1 reasoning is solved by the minimization of K-L information under the marginal distribution constraints. We showed ISP can be applied to Type 1 probabilistic reasoning and proposed an efficient propagation algorithm on Junction Trees. Both the time and the space complexities of the proposed algorithm are lower than those of the previous algorithms using ISP.

Appendix: Proof of Theorem 2

First we show the following lemma for proving Theorem 2. $P_1^l(X_1, \ldots, X_n)$ denotes the $P(X_1, \ldots, X_n)$ calculated at the lth loop [7] in Procedure 1. $P_2^h(N)$ denotes the $P(N)$ that is calculated in Procedure 2 after messages have passed through $h$ RINs. We assume that the order of the RINs are $X_1, \ldots, X_k$.

Lemma 3 *When a message reaches a clique N at $h = k$ in Procedure 2, the calculated distribution of the clique N satisfies the following equation.*

$$P_2^k(N) = \sum_{X \notin N} P_1^k(X_1, \ldots, X_n). \tag{20}$$

Proof:

We prove Lemma 3 by the inductive method. It is obvious at $h = 0$.

We assume that Lemma 3 holds in a clique node N_u at $h = k$.

1) In the case where the next intersection node is a RIN

Let the next intersection node be $D_{u,v}$. From the assumption,

$$\begin{aligned} \sum_{X \notin D_{u,v}} P_2^k(N_v) &= \sum_{X \notin D_{u,v}} P_1^k(X_1, \ldots, X_n) \\ &= P_1^k(D_{u,v}). \end{aligned} \tag{21}$$

Remark $N_u = N_v$, because $D_{u,v}$ is a RIN.

Following Procedure 2, $P_2^{k+1}(N_v)$ is calculated as follows:

$$P_2^{k+1}(N_v) = P_1^k(N_v) \frac{P^*(D_{u,v})}{P_1^k(D_{u,v})}. \tag{22}$$

[7] The index $i = l$ at the lth loop in Procedure 1.

Procedure 1 calculates $P_1^{k+1}(X_1,\ldots,X_n)$ as follows:

$$\begin{aligned} & P_1^{k+1}(X_1,\ldots,X_n) \\ &= P_1^k(X_1,\ldots,X_n)\frac{P^*(D_{u,v})}{P_1^k(D_{u,v})}. \end{aligned} \tag{23}$$

From the definition of RTs, the marginal distribution $P_1^{k+1}(N_v)$ is given as follows:

$$\begin{aligned} & \sum_{X\notin N_v} P_1^{k+1}(X_1,\ldots,X_n) \\ &= \sum_{X\notin N_v} P_1^k(X_1,\ldots,X_n)\frac{P^*(D_{u,v})}{P_1^k(D_{u,v})} \\ &= P_1^k(N_v)\frac{P^*(D_{u,v})}{P_1^k(D_{u,v})}. \end{aligned} \tag{24}$$

Thus, from Formulas (22), (24), Formula (20) holds in N_v at $h=k+1$ as follows:

$$P_2^{k+1}(N_v) = \sum_{X\notin N_v} P_1^{k+1}(X_1,\ldots,X_n). \tag{25}$$

2) In the case where the next intersection node is not a RIN

Let the next intersection node be $D_{u,v}$. RT_u denotes the maximum subtree of the RT that includes N_u and not N_v. Let S_{N_u} be the vector or the set of random variables in RT_u and S_{N_v} the vector or the set of random variables in the complement tree of RT_u.

By using the intersection node $D_{u,v}$, the joint distribution calculated at the kth loop in Procedure 1 can be represented as follows:

$$P_1^k(X_1,\ldots,X_n) = \frac{P_1^k(S_{N_u})P_1^{k-r}(S_{N_v})}{P_1^{k-r}(D_{u,v})}, \tag{26}$$

where r is the number of the clique nodes in RT_u.

The marginal distribution of N_v with respect to the joint distribution of Formula(26) is given by

$$\sum_{X\notin N_v} P_1^k(X_1,\ldots,X_n) = \frac{P_1^{k-r}(N_v)P_1^k(N_u)}{P_1^{k-r}(D_{u,v})}. \tag{27}$$

From the assumption and the calculation process of Procedure 2,

$$\begin{aligned} P_2^k(N_u) &= \sum_{X\notin N_u} P_1^k(X_1,\ldots,X_n) = P_1^k(N_u), \\ P_2^{k-r}(N_v) &= \sum_{X\notin N_v} P_1^{k-r}(X_1,\ldots,X_n) = P_1^{k-r}(N_v). \end{aligned}$$

Following Procedure 2, $P_2^k(N_v)$ is calculated as follows:

$$\begin{aligned} P_2^k(N_v) &= P_2^{k-r}(N_v)\frac{P_2^k(D_{u,v})}{\sum_{X\notin D_{u,v}} P_2^{k-r}(N_v)} \\ &= P_1^{k-r}(N_v)\frac{P_1^k(D_{u,v})}{P_1^{k-r}(D_{u,v})}. \end{aligned} \tag{28}$$

Thus, from Formulas (27), (28), Formula (20) holds in N_v, which is the next clique of N_u, at $h=k$ as follows:

$$P_2^k(N_v) = \sum_{X\notin N_v} P_1^k(X_1,\ldots,X_n). \tag{29}$$

Thus, Theorem 2 can be proved from Lemma 2 and Lemma 3.

Acknowledgements

This research was supported in part by the Ministry of Education under Grant-Aids (c) No.15560338 for Scientific Research.

References

[Aji and Mcliece 2000] S.M. Aji, R.J. Mcliece, *The Generalized Distributive Law*, IEEE Trans. IT, Vol.46 No.2, 2000.

[Csiszar 1975] I. Csiszar, *I-divergence geometry of probability distributions and minimization problems*, The Annals of Probability, Vol. 13, No. 1, 146-158, 1975.

[Ireland and Kullback 1968] C. T. Ireland and S. Kullback, *Contingency tables with given marginals*, Biometrika, Vol. 55, 179-188, 1968.

[Jensen 1996] F. V. Jensen, *An introduction to Bayesian networks*, University College London Press, London, 1996.

[Jirousek and Preucil 1995] R. Jirousek and S. Preucil, *On the effective implementation of the iterative proportional fitting prodecure*, Computational Statistics and Data Analysis, North-Holland, 1995.

[Kschischang *et al.* 2001] F.R. Kschischang, B.J. Fey and H. Loeliger, *Factor Graphs and the Sum-Product Algorithm*, IEEE Trans. IT, Vol.47 No.2, 2001.

[Kullback 1959] S. Kullback, *Information Theory and Statistics*, Wiley, New York, 1959.

[Matsushima *et al.* 2001] T. Matsushima, T.K. Matsushima and S. Hirasawa *An Iterative Algorithm for Calculating Posterior Probability and Model Representation* , Proceedings of IEEE Int. Symp. on Information Theory, 2001.

[Matsushima *et al.* 2002A] T. Matsushima, T.K. Matsushima and S. Hirasawa *An Alternative Algorithm for Calculating Posterior Probability and Decoding*, Proceedings of IEEE Int. Symp. on Information Theory, 2002.

[Matsushima *et al.* 2002B] T. Matsushima, T.K. Matsushima and S. Hirasawa *Calculation of Generalized Posterior Distribution on Junction Graphs*, Proceedings of the 25th Symposium on Information Theory and Its Applications, 2002.

[McElice *et al.* 1998] R.J. McElice, D.J.C. MacKay and J. Cheng, *Turbo decoding as an instance of Pearl's "Belief Propagation"*, IEEE J. Sel. Areas Commun., Vol.16 No.2, 1998.

[Pearl 1988] J. Pearl, *Probabilistic reasoning in intelligent systems* Morgan Kaufmann, 1988.

[Pearl 1990] J. Pearl *Jeffry's rule, passage of experiments and Neo-Bayesianism*, Knowledge Representation and Defeasible Reasoning, 245-265, Kluwer Academic Publisher, 1990.

[Skyrms1985] B. Skyrms, *Maximum Entropy Inference as a Special Case of Conditionalization*, Synthese, 63, 1985.

[Valtorta *et al.* 2002] M. Valtorta, Y. Kim, J. Vomlel, *Soft Evidential Update for Probabilistic Multiagent Systems*, International Journal of Approximate Reasoniong, 29, 1, 2002.

[Wen 1990] W.X. Wen, *Minimum Cross Entropy Reasoning in Recursive Causal Networks*, Uncertainty in artificial Intelligence 4, 105-119, 1990.