Incorporating Uncertain Evidence Into Arithmetic Circuits Representing Probability Distributions

Hei Chan HEI@IST.HOKUDAI.AC.JP
Graduate School of Information Science and Technology
Hokkaido University
Sapporo (Japan)

Abstract

Arithmetic circuits have been used as tractable representations of probability distributions, either generated from models such as Bayesian networks, sum-product networks and Probability Sentential Decision Diagrams, or directly from data. An interesting question is how we can incorporate uncertain evidence, which specifies that the marginal probabilities of a variable has to undergo certain changes, directly into an arithmetic circuit and then perform reasoning on it to compute the probability distribution after incorporating this uncertain evidence. In this paper, we show that we can incorporate uncertain evidence on a variable by setting indicators of this variable in the arithmetic circuit to non-negative values based on the likelihood ratios in Pearl's method of virtual evidence and the current marginal probabilities of this variable. For tractable computation of these marginal probabilities, the arithmetic circuit has to satisfy the properties of decomposability and smoothness, and we show that an algorithm using a downward pass can compute these marginal probabilities for all single variables. We show a procedure of how to incorporate virtual evidence, including multiple pieces of virtual evidence.

Keywords: Arithmetic circuit; Bayesian network; Belief revision; Uncertain evidence; Virtual evidence.

1. Introduction

Arithmetic circuits (AC) have been used recently to represent probability distributions in a tractable form. These probability distributions may be generated by models such as Bayesian networks (BN) (Pearl, 1988), sum-product networks (SPN) (Poon and Domingos, 2011) and Probability Sentential Decision Diagrams (PSDD) (Kisa et al., 2014), or directly from data (Lowd and Domingos, 2008). After constructing the arithmetic circuit to represent the probability distribution, we can perform inference on it to find the answers to various probability queries given different pieces of evidence.

In the problem of belief revision, beliefs represented by probability distributions may be revised due to evidence received from different sources. This includes uncertain evidence, where for some variable, instead of being certain of taking a value, its marginal probabilities must satisfy the constraints as posed by this uncertain evidence. Incorporating uncertain evidence may be based on different methods, including soft evidence (Jeffrey, 1990) and virtual evidence (Pearl, 1988), although it has been shown both methods are based on the principle of probability kinematics, and thus can be translated between each other (Chan and Darwiche, 2005).

The problem we address here is whether belief revision based on uncertain evidence can be directly applied to an arithmetic circuit representing a probability distribution. The answer is yes, if we set values of indicators to non-negative values other than 0 or 1 (in all previous literature, values of indicators are set to only 0 or 1 based on whether the indicator is consistent with the given evidence). To compute the values of the indicators that we must set to, we have to know the strength of the uncertain evidence (for example, in virtual evidence, the strength is given by likelihood ratios) and the marginal probabilities of the variable where uncertain evidence is obtained. It has been proved previously that marginal probabilities can be computed by an arithmetic circuit that satisfies the properties of decomposability and smoothness (Choi and Darwiche, 2017). In this paper, we also show that they can be computed in such arithmetic circuit by a downward pass that computes partial derivatives (Darwiche, 2003). As the operations of upward pass and downward pass can be done in time linear in the number of nodes in the arithmetic circuit, our procedure of incorporating uncertain evidence is tractable.

Our paper is structured as follows. We first define the preliminaries including factors and distributions. We next show how belief revision given uncertain evidence (in our case, virtual evidence) can be incorporated into a probability distribution. We then discuss how arithmetic circuits can be used to represent a probability distribution and how we can compute marginal probabilities for an arithmetic circuit that is both decomposable and smooth. We finally show our procedure of incorporating uncertain evidence into an arithmetic circuit that is both decomposable and smooth.

2. Preliminaries

In this paper, upper-case letters X are used to denote variables and lower-case letters x are used to denote possible values of X. The variable X is discrete and we assume may take on m_X possible values, which we denote as non-negative integers $x \in \{0, \ldots, m_X - 1\}$. Bold letters $\mathbf{X}$ and $\mathbf{x}$ are used to denote sets of variables and their possible instantiations respectively.

We call the assignment of a variable X to any of its possible values x, denoted as $X = x$, as *evidence* on X. An evidence $X = x$ is *consistent* with instantiation $\mathbf{y}$, denoted as $x \sim \mathbf{y}$, if $X \in \mathbf{Y}$ and is assigned to x in $\mathbf{y}$ or $X \notin \mathbf{Y}$; otherwise, we denote $x \nsim \mathbf{y}$. In this paper, we will use the function $\mathbf{1}(x \sim \mathbf{y})$, where $\mathbf{1}(x \sim \mathbf{y}) = 1$ if $x \sim \mathbf{y}$ and $\mathbf{1}(x \sim \mathbf{y}) = 0$ if $x \nsim \mathbf{y}$.

We now define the notions of *factor* and *distribution*, which can be most easily seen as tables of values for each instantiation $\mathbf{x}$ of $\mathbf{X}$.

Definition 1 *A factor $f(\mathbf{X})$ over variables $\mathbf{X}$ maps each instantiation $\mathbf{x}$ into a non-negative number $f(\mathbf{x})$. A factor is a distribution if $\sum_{\mathbf{X}=\mathbf{x}} f(\mathbf{x}) = 1$.*

The following operations can be applied to a factor or a distribution:

- *Normalization*: A distribution $Pr(\mathbf{X})$ is said to be obtained from factor $f(\mathbf{X})$ by normalization if $Pr(\mathbf{x}) = f(\mathbf{x}) / \sum_{\mathbf{X}=\mathbf{x}} f(\mathbf{x})$ for each instantiation $\mathbf{x}$ of $\mathbf{X}$, where the sum $\sum_{\mathbf{X}=\mathbf{x}} f(\mathbf{x})$ is called the *normalization constant*.
- *Marginalization*: Given variable $X \in \mathbf{X}$ and $\mathbf{Y} = \mathbf{X} \setminus \{X\}$, the factor $(\sum_X f)(\mathbf{Y})$ is said to be obtained from factor $f(\mathbf{X})$ by marginalization over X if $(\sum_X f)(\mathbf{y}) =$

A	B	$Pr^{\alpha}(A,B)$	$Pr^{\beta}(A,B)$
0	0	0.1	0.24
0	1	0.2	0.48
1	0	0.3	0.12
1	1	0.4	0.16

Figure 1: Distributions over binary variables A, B. $Pr^{\beta}(A,B)$ is obtained from $Pr^{\alpha}(A,B)$ based on probability kinematics over A.

A	B	Z	$Pr^{\gamma}(A,B,Z)$	$Pr^{\gamma}(A,B,Z \mid Z=1)$
0	0	0	0.04	0
0	0	1	0.06	0.24
0	1	0	0.08	0
0	1	1	0.12	0.48
1	0	0	0.27	0
1	0	1	0.03	0.12
1	1	0	0.36	0
1	1	1	0.04	0.16

Figure 2: Distributions over binary variables A, B, Z. $Pr^{\gamma}(A,B,Z)$ is constructed from $Pr^{\alpha}(A,B)$ based on virtual evidence $Z=1$ with likelihood ratios $L(A=0 \mid Z=1) : \lambda(A=1 \mid Z=1) = 6:1$. Note that $Pr^{\alpha}(A,B)$ and $Pr^{\beta}(A,B)$ in Figure 1 can be obtained from $Pr^{\gamma}(A,B,Z)$ and $Pr^{\gamma}(A,B,Z \mid Z=1)$ respectively by marginalization over Z.

$\sum_{X=x} f(x,\mathbf{y})$ for each instantiation $\mathbf{y}$ of $\mathbf{Y}$, and for simplicity, we denote this as $f(\mathbf{Y}) \equiv (\sum_X f)(\mathbf{Y})$. Marginalization is commutative, and thus can be applied over a set of variables one by one in any order, and call the resulting factor a *factor marginal*.

- *Conditioning*: Given variable $X \in \mathbf{X}$ and evidence $X = x$, the distribution $Pr(\mathbf{X} \mid X = x)$ is said to be obtained from conditioning on $X = x$ by first constructing a new factor $f(\mathbf{X} \mid X = x)$ where $f(\mathbf{x} \mid X = x) = \mathbf{1}(x \sim \mathbf{x})Pr(\mathbf{x})$, and then applying normalization on $f(\mathbf{X} \mid X = x)$ to obtain the distribution $Pr(\mathbf{X} \mid X = x)$. $Pr(\mathbf{X} \mid X = x)$ is usually applied marginalization over X, which is trivial since only values $Pr(\mathbf{x} \mid X = x)$ where $x \sim \mathbf{x}$ may be non-zero. Conditioning is commutative, and thus can be applied over a set of evidence one by one in any order, and call the resulting distribution a *conditional distribution*.

3. Uncertain evidence

Given distribution $Pr(\mathbf{X})$, the notion of having *uncertain evidence* on variable $X \in \mathbf{X}$ is that the marginal probabilities of X have to undergo certain changes in order to incorporate

this uncertain evidence. There are different ways to specify uncertain evidence. One such method is *soft evidence* (Jeffrey, 1990), which specifies the marginal probabilities of the variable X after the uncertain evidence is incorporated. Another method is to use *virtual evidence* (Pearl, 1988), which recasts this uncertain evidence as evidence on some new virtual variable and specifies the likelihood ratios of this evidence given each value of the variable X. Both methods are based on *probability kinematics*, which enforces the *rigidity condition* on the new distribution after incorporating the uncertain evidence.

Definition 2 *Given two distributions $Pr(\mathbf{X})$ and $Pr'(\mathbf{X})$, $Pr'(\mathbf{X})$ is said to be obtained from $Pr(\mathbf{X})$ based on probability kinematics over variable X if for every value x of X and every instantiation $\mathbf{y}$ of variables $\mathbf{Y} \subseteq \mathbf{X}$, $Pr(\mathbf{y} \mid X = x) = Pr'(\mathbf{y} \mid X = x)$ (Jeffrey, 1990).*

For example, two distributions $Pr^\alpha(A, B)$ and $Pr^\beta(A, B)$ are shown in Figure 1. The marginal probabilities of A are $(Pr^\alpha(A = 0), Pr^\alpha(A = 1)) = (0.3, 0.7)$ and $(Pr^\beta(A = 0), Pr^\beta(A = 1)) = (0.72, 0.28)$ respectively. We can check that $Pr^\beta(A, B)$ is obtained from $Pr^\alpha(A, B)$ based on probability kinematics over A. For example, $(Pr^\alpha(B = 0 \mid A = 0), Pr^\alpha(B = 1 \mid A = 0)) = (Pr^\beta(B = 0 \mid A = 0), Pr^\beta(B = 1 \mid A = 0)) = (1/3, 2/3)$ and $(Pr^\alpha(B = 0 \mid A = 1), Pr^\alpha(B = 1 \mid A = 1)) = (Pr^\beta(B = 0 \mid A = 1), Pr^\beta(B = 1 \mid A = 1)) = (3/7, 4/7)$.

One main difference between soft evidence and virtual evidence is that the latter is commutative while the former is not (Wagner, 2002). This means that the order in which we incorporate the uncertain evidence does not matter if we specify them using virtual evidence. For this reason, we use virtual evidence in this paper, which we define next. For a review of the two methods and how we can translate between them, see Chan and Darwiche (2005).

Definition 3 *Given distribution $Pr(\mathbf{X})$ and variable $X \in \mathbf{X}$ with possible values $x \in \{0, \ldots, m_X - 1\}$, virtual evidence on X is specified in the form of evidence on a new variable $Z = 1$ (we always assume Z is binary with possible values $\{0, 1\}$), and non-negative likelihood ratios $L(X) = (L(X = 0), \ldots, L(X = m_X - 1))$ such that $Pr(Z = 1 \mid X = 0) : \ldots : Pr(Z = 1 \mid X = m_X - 1) = L(X = 0) : \ldots : L(X = m_X - 1)$, i.e., there is some positive number k such that for all $x \in \{0, \ldots, m_X - 1\}$, $Pr(Z = 1 \mid X = x) = k \cdot L(X = x)$. Moreover, the evidence $Z = 1$ is independent of variables $\mathbf{Y} \subseteq \mathbf{X}$ given X, i.e., $Pr(Z = 1 \mid X = x, \mathbf{Y} = \mathbf{y}) = Pr(Z = 1 \mid X = x)$ for all values x of X and instantiations $\mathbf{y}$ of $\mathbf{Y}$. The distribution after incorporating the virtual evidence is the one obtained by conditioning on $Z = 1$, i.e., $Pr(\mathbf{X} \mid Z = 1)$ (Pearl, 1988).*

For example, given distribution $Pr^\alpha(A, B)$ in Figure 1, we need to incorporate virtual evidence on A which is specified in the form of evidence on a new variable $Z = 1$, and likelihood ratios $L(A) = (L(A = 0), L(A = 1)) = (6, 1)$. Based on this virtual evidence, distribution $Pr^\gamma(A, B, Z)$, shown in Figure 2, is constructed, from which we can apply marginalization over Z to obtain $Pr^\alpha(A, B)$, shown in Figure 1. We can see that $Pr^\gamma(Z = 1 \mid A = 0) = 0.6$ and $Pr^\gamma(Z = 1 \mid A = 1) = 0.1$, satisfying the likelihood ratios $Pr(Z = 1 \mid A = 0) : Pr(Z = 1 \mid A = 1) = L(A = 0) : L(A = 1) = 6 : 1$, and $Pr^\gamma(Z = 1 \mid A = 0, B = 0) = Pr^\gamma(Z = 1 \mid A = 0, B = 1) = Pr^\gamma(Z = 1 \mid A = 0)$ and $Pr^\gamma(Z = 1 \mid A = 1, B = 0) = Pr^\gamma(Z = 1 \mid A = 1, B = 1) = Pr^\gamma(Z = 1 \mid A = 1)$, satisfying the independence

condition. From $Pr^{\gamma}(A, B, Z)$, if we apply conditioning on evidence $Z = 1$, we obtain $Pr^{\gamma}(A, B, Z \mid Z = 1)$, and after applying marginalization over Z (which is trivial), we obtain $Pr^{\beta}(A, B)$, shown in Figure 1, which is the new distribution after incorporating the virtual evidence on A.

It has been shown that given virtual evidence on X which satisfies the likelihood ratios and the independence condition, there is a unique distribution after incorporating the virtual evidence, and it has also been shown that this distribution is obtained from the original distribution based on probability kinematics on X (Chan and Darwiche, 2005).

Corollary 4 *Given distribution $Pr(\mathbf{X})$ and virtual evidence $Z = 1$ on $X \in \mathbf{X}$ specified by Definition 3, the probability of instantiation $\mathbf{x}$ of $\mathbf{X}$ after incorporating the virtual evidence, if $X = x \sim \mathbf{x}$, is given by:*

$$Pr(\mathbf{x} \mid Z = 1) = \frac{L(X = x)Pr(\mathbf{x})}{\sum_{X=x' \in \{0,\ldots,m_X-1\}} L(X = x')Pr(X = x')} \tag{1}$$

Incorporating virtual evidence in Equation 1 can be seen as first constructing a new factor $f(\mathbf{X} \mid Z = 1)$ where $f(\mathbf{x} \mid Z = 1) = L(X = x)Pr(\mathbf{x})$, and then applying normalization on $f(\mathbf{X} \mid Z = 1)$ to obtain the distribution $Pr(\mathbf{X} \mid Z = 1)$ (we can easily show the normalization constant is the one given in Equation 1).

For any virtual evidence on X, only the likelihood ratios $L(X = 0) : \ldots : L(X = m_X - 1)$ matter but not the individual values $L(X = x)$. In fact, given $L(X) = (L(X = 0), \ldots, L(X = m_X - 1))$ and $Pr(X) = (Pr(X = 0), \ldots, Pr(X = m_X - 1))$, we can compute $L^1 = (L^1(X = 0), \ldots, L^1(X = m_X - 1))$ which has the same likelihood ratios as L and the normalization constant (from Equation 1) is equal to 1.

Corollary 5 *Given $L(X) = (L(X = 0), \ldots, L(X = m_X - 1))$ and $Pr(X) = (Pr(X = 0), \ldots, Pr(X = m_X - 1))$, we compute $L^1 = (L^1(X = 0), \ldots, L^1(X = m_X - 1))$ as, for all $x \in \{0, \ldots, m_X - 1\}$:*

$$L^1(X = x) = \frac{L(X = x)}{\sum_{X=x' \in \{0,\ldots,m_X-1\}} L(X = x')Pr(X = x')} \tag{2}$$

Since we have $L^1(X = 0) : \ldots : L^1(X = m_X - 1) = L(X = 0) : \ldots : L(X = m_X - 1)$ and $\sum_{X=x' \in \{0,\ldots,m_X-1\}} L^1(X = x')Pr(X = x') = 1$, the probability of instantiation $\mathbf{x}$ of $\mathbf{X}$, if $X = x \sim \mathbf{x}$, is given by (from Equation 1):

$$Pr(\mathbf{x} \mid Z = 1) = L^1(X = x)Pr(\mathbf{x}) \tag{3}$$

For example, when incorporating virtual evidence in Figure 2, since we have $L(A) = (L(A = 0), L(A = 1)) = (6, 1)$ and $Pr^{\gamma}(A) = (Pr^{\gamma}(A = 0), Pr^{\gamma}(A = 1)) = (0.3, 0.7)$, we can compute $L^1(A) = (L^1(A = 0), L^1(A = 1)) = (6/(6 * 0.3 + 1 * 0.7), 1/(6 * 0.3 + 1 * 0.7)) = (2.4, 0.4)$. We can easily verify this by checking Figure 1, with $Pr^{\beta}(A = 0, B = 0) = L^1(A = 0)Pr^{\alpha}(A = 0, B = 0) = 2.4 * 0.1 = 0.24$, $Pr^{\beta}(A = 0, B = 1) = L^1(A = 0)Pr^{\alpha}(A = 0, B = 1) = 2.4 * 0.2 = 0.48$, $Pr^{\beta}(A = 1, B = 0) = L^1(A = 1)Pr^{\alpha}(A = 1, B = 0) = 0.4 * 0.3 = 0.12$, and $Pr^{\beta}(A = 1, B = 1) = L^1(A = 1)Pr^{\alpha}(A = 1, B = 1) = 0.4 * 0.4 = 0.16$.

The following procedure shows how virtual evidence can be incorporated into a Bayesian network. For a review of how uncertain evidence can be represented in Bayesian networks in general, see Ben Mrad et al. (2015).

Corollary 6 *In Bayesian networks, the following procedure allows us to compute the probability values after incorporating virtual evidence on X (Pearl, 1988):*

1. *Adding a new node Z as a child of X.*
2. *Specifying the conditional probability table of Z given X such as that $Pr(Z = 1 \mid X = 0) : \ldots : Pr(Z = 1 \mid X = m_X - 1) = L(X = 0) : \ldots : L(X = m_X - 1)$.*
3. *Conditioning on evidence $Z = 1$.*

4. Arithmetic circuits

An *arithmetic circuit* (AC) is a structure that can be used to represent a factor $f(\mathbf{X})$ (Darwiche, 2003). The input of the arithmetic circuit is denoted as Λ, which we call *indicators*.

Definition 7 *An arithmetic circuit $\mathcal{AC}$ over variables $\mathbf{X}$ is a rooted directed acyclic graph whose internal nodes are either addition nodes labeled with + or multiplication nodes labeled with *, and whose leaf nodes are either indicators $\lambda_{X=x} \in \Lambda$ where x is a value of variable $X \in \mathbf{X}$, or parameters θ, which both must be non-negative values. For every node n in $\mathcal{AC}$, we denote $vars(n)$ as all variables $X \in \mathbf{X}$ where some indicator $\lambda_{X=x}$ appears at or under node n.*

To compute the value of the marginal $f(\mathbf{y})$ using an arithmetic circuit, we first have to compute the corresponding $\Lambda(\mathbf{y})$, such that the values of each indicator are set to 0 or 1 based on whether the indicator is consistent with $\mathbf{y}$.

Definition 8 *Given instantiation $\mathbf{y}$ of $\mathbf{Y} \subseteq \mathbf{X}$, the corresponding indicators, which we denote as $\Lambda(\mathbf{y})$, are assigned as $\lambda_{X=x} \leftarrow \mathbf{1}(x \sim \mathbf{y})$, for all values x of all variables $X \in \mathbf{X}$.*

After setting the input of the arithmetic circuit $\mathcal{AC}$ as $\Lambda(\mathbf{y})$, we can compute the output of the arithmetic circuit $\mathcal{AC}(\Lambda(\mathbf{y}))$ by evaluating it using an upward pass (visiting children before parents) according to the operations (+ or *) of the internal nodes (Darwiche, 2003), and returning the value of the root node as $\mathcal{AC}(\Lambda(\mathbf{y}))$.

Given a factor $f(\mathbf{X})$, we say that $\mathcal{AC}$ *computes factor* $f(\mathbf{X})$ if $\mathcal{AC}(\Lambda(\mathbf{x})) = \mathbf{f}(\mathbf{x})$ for all full instantiations $\mathbf{x}$ of $\mathbf{X}$, and that $\mathcal{AC}$ *computes marginals of factor* $f(\mathbf{X})$ if $\mathcal{AC}(\Lambda(\mathbf{y})) = \mathbf{f}(\mathbf{y})$ for all partial instantiations $\mathbf{y}$ of $\mathbf{Y} \subseteq \mathbf{X}$. While an arithmetic circuit that computes the marginals of a factor also computes the factor, an arithmetic circuit that computes a factor does not necessarily compute its marginals. It has been shown that an arithmetic circuit must obey two properties for it to be able to compute marginals correctly: *decomposability* and *smoothness*.

Definition 9 *An arithmetic circuit is decomposable iff for every *-node n, $vars(c_1) \cap vars(c_2) = \emptyset$ for every pair of children c_1 and c_2 (Darwiche, 2001).*

Definition 10 *An arithmetic circuit is smooth iff it contains at least one indicator for each variable $X \in \mathbf{X}$, and for every +-node n, $vars(c) = vars(n)$ for every child c (Darwiche, 2001).*

Corollary 11 *If an arithmetic circuit $\mathcal{AC}$ that computes factor $f(\mathbf{X})$ is both decomposable and smooth, then it also computes the marginals of factor $f(\mathbf{X})$ (Choi and Darwiche, 2017).*

For example, an arithmetic circuit that is both decomposable and smooth can be generated from a Bayesian network (Darwiche, 2003), and thus can be used to compute marginals.

Finally we show one more property of an arithmetic circuit that is both decomposable and smooth. After computing the marginals of the factor $f(\mathbf{y})$ by evaluating $\mathcal{AC}(\mathbf{y})$ using an upward pass, we can differentiate it using a downward pass (visiting parents before children) according to the operations (+ or *) of the internal nodes (Darwiche, 2003); for full algorithm, see (Darwiche, 2009, Page 293, Algorithm 34). This differentiation computes the partial derivative $\partial\mathcal{AC}(\Lambda(\mathbf{y}))/\partial\lambda_{X=x}$ for each indicator $\lambda_{X=x}$. We can show that given an arithmetic circuit that is both decomposable and smooth, this partial derivative is equal to $f(\mathbf{y}, x)/\lambda_{X=x}$ if $X \notin \mathbf{Y}$. The equivalent result was proved for arithmetic circuits representing Bayesian networks (Darwiche, 2003), but we believe this result for general arithmetic circuits is original. A proof sketch is shown in the Appendix.

Theorem 12 *If an arithmetic circuit $\mathcal{AC}$ that computes factor $f(\mathbf{X})$ is both decomposable and smooth, then for any value x of X and instantiation $\mathbf{y}$ of $\mathbf{Y}$ where $X \notin \mathbf{Y} \subseteq \mathbf{X}$, we have $\partial\mathcal{AC}(\Lambda(\mathbf{y}))/\partial\lambda_{X=x} = f(\mathbf{y}, x)/\lambda_{X=x}$.*

For example, if no evidence has been set, i.e., indicator $\Lambda(\top)$ is set as $\lambda_{X=x} \leftarrow 1$ for all values x of all variables $X \in \mathbf{X}$, the marginal $f(x)$ for some value x of variable $X \in \mathbf{X}$ can be computed as $\partial\mathcal{AC}(\Lambda(\top))/\partial\lambda_{X=x} = f(x)/\lambda_{X=x} = f(x)$.

5. Incorporating uncertain evidence into arithmetic circuits

We showed in Equation 3 that the distribution after incorporating virtual evidence on X can be computed if we know the marginal distribution $Pr(X)$, and in Theorem 12 that this marginal distribution can be computed by an upward pass followed by a downward pass of an arithmetic circuit that is both decomposable and smooth. Therefore, we can develop a procedure where we can directly incorporate virtual evidence into an arithmetic circuit that is both decomposable and smooth, by allowing the indicators to take non-negative values other than 0 or 1. A proof sketch is shown in the Appendix.

Theorem 13 *Given variables $\mathbf{X} = \{X_1, \ldots, X_n\}$, virtual evidence $Z = 1$ on $X \in \mathbf{X}$ specified by Definition 3, and an arithmetic circuit $\mathcal{AC}$ which is both decomposable and smooth and computes distribution $Pr(\mathbf{X})$, the following procedure allows us to compute marginal values after incorporating the virtual evidence, for example, the marginal value of instantiation $\mathbf{y}$ of $\mathbf{Y} \subseteq \mathbf{X}$, i.e., $Pr(\mathbf{y} \mid Z = 1)$, and also $Pr(\mathbf{y}, X_i = x_i \mid Z = 1)$ for all values x_i of all variables $X_i \notin \mathbf{Y}$:*

1. *Set values of the indicators $\Lambda^0(\top)$ as $\lambda_{X_i=x_i} \leftarrow 1$ for all values x_i of all variables $X_i \in \mathbf{X}$ (no evidence has been set).*
2. *Use an upward pass to compute $\mathcal{AC}(\Lambda^0(\top)) = Pr(\top) = 1$.*
3. *Use a downward pass to compute $\partial\mathcal{AC}(\Lambda^0(\top))/\partial\lambda_{X_i=x_i} = Pr(X_i = x_i)/\lambda_{X_i=x_i} = Pr(X_i = x_i)$ for all values x_i of all variables $X_i \in \mathbf{X}$ (from Theorem 12).*

4. *Given* $L(X) = (L(X = 0), \ldots, L(X = m_X - 1))$ *(from virtual evidence) and* $Pr(X) = (Pr(X = 0), \ldots, Pr(X = m_X - 1))$ *(from Step 3), compute* $L^1(X) = (L^1(X = 0), \ldots, L^1(X = m_X - 1))$ *as, for all* $x \in \{0, \ldots, m_X - 1\}$ *(from Equation 2):*

$$L^1(X = x) = \frac{L(X = x)}{\sum_{X=x' \in \{0,\ldots,m_X-1\}} L(X = x')Pr(X = x')} \tag{4}$$

5. *Set values of the indicators* $\Lambda^1(\mathbf{y})$ *as:*
   - $\lambda_{X=x} \leftarrow \mathbf{1}(x \sim \mathbf{y})L^1(X = x)$ *for all values* x *of* X*.*
   - *For all* $X_i \neq X$, $\lambda_{X_i=x_i} \leftarrow \mathbf{1}(x_i \sim \mathbf{y})$ *for all values* x_i *of* X_i*.*
6. *Use an upward pass to compute* $\mathcal{AC}(\Lambda^1(\mathbf{y})) = Pr(\mathbf{y} \mid Z = 1)$*.*
7. *Use a downward pass to compute* $\partial\mathcal{AC}(\Lambda^1(\mathbf{y}))/\partial\lambda_{X_i=x_i} = Pr(\mathbf{y}, X_i = x_i \mid Z = 1)/\lambda_{X_i=x_i}$ *for all values* x_i *of all variables* $X_i \notin \mathbf{Y}$*.*

For example, in Figure 1, $Pr^\alpha(A, B)$ can be expressed by the arithmetic circuit $\mathcal{AC}(\Lambda) = 0.4 * \lambda_{B=0} * (1/4 * \lambda_{A=0} + 3/4 * \lambda_{A=1}) + 0.6 * \lambda_{B=1} * (1/3 * \lambda_{A=0} + 2/3 * \lambda_{A=1})$, which is both decomposable and smooth. The partial derivatives are given by $\partial\mathcal{AC}(\Lambda)/\partial\lambda_{A=0} = 0.1 * \lambda_{B=0} + 0.2 * \lambda_{B=1}$, $\partial\mathcal{AC}(\Lambda)/\partial\lambda_{A=1} = 0.3 * \lambda_{B=0} + 0.4 * \lambda_{B=1}$, $\partial\mathcal{AC}(\Lambda)/\partial\lambda_{B=0} = 0.1 * \lambda_{A=0} + 0.3 * \lambda_{A=1}$, and $\partial\mathcal{AC}(\Lambda)/\partial\lambda_{B=1} = 0.2 * \lambda_{A=0} + 0.4 * \lambda_{A=1}$. Assuming we have virtual evidence on A in the form of $Z = 1$, and $L(A) = (L(A = 0), L(A = 1)) = (6, 1)$, we now follow the procedure in Theorem 13 to compute $Pr^\alpha(B = 1 \mid Z = 1)$ and $Pr^\alpha(A = 0 \mid Z = 1)$ after incorporating this virtual evidence:

1. Set values of the indicators $\Lambda^0(\top)$ as $\lambda_{A=0} \leftarrow 1$, $\lambda_{A=1} \leftarrow 1$, $\lambda_{B=0} \leftarrow 1$, and $\lambda_{B=1} \leftarrow 1$ (no evidence has been set).
2. Use an upward pass to compute $\mathcal{AC}(\Lambda^0(\top)) = Pr^\alpha(\top) = 1$.
3. Use a downward pass to compute $\partial\mathcal{AC}(\Lambda^0(\top))/\partial\lambda_{A=0} = Pr^\alpha(A = 0) = 0.3$ and $\partial\mathcal{AC}(\Lambda^0(\top))/\partial\lambda_{A=1} = Pr^\alpha(A = 1) = 0.7$.
4. Given $L(A) = (6, 1)$ and $Pr(A) = (0.3, 0.7)$, compute $L^1(A) = (6/(6 * 0.3 + 1 * 0.7), 1/(6 * 0.3 + 1 * 0.7)) = (2.4, 0.4)$.
5. Set values of the indicators $\Lambda^1(B = 1)$ as $\lambda_{A=0} \leftarrow L^1(A = 0) = 2.4$, $\lambda_{A=1} \leftarrow L^1(A = 1) = 0.4$, $\lambda_{B=0} \leftarrow 0$, and $\lambda_{B=1} \leftarrow 1$.
6. Use an upward pass to compute $\mathcal{AC}(\Lambda^1(B = 1)) = 0.4*0*(1/4*2.4+3/4*0.4)+0.6*1*(1/3*2.4+2/3*0.4) = 0.48+0.16 = 0.64$ (same as computed by $Pr^\beta(B = 1 \mid Z = 1)$ in Figure 1).
7. Use downward pass to compute $\partial\mathcal{AC}(\Lambda^1(B = 1))/\partial\lambda_{A=0} = 0.1 * \lambda_{B=0} + 0.2 * \lambda_{B=1} = 0.2 = Pr(A = 0, B = 1 \mid Z = 1)/\lambda_{A=0}$, which means $Pr(A = 0, B = 1 \mid Z = 1) = 0.2 * \lambda_{A=0} = 0.2 * 2.4 = 0.48$ (same as computed by $Pr^\beta(A = 0, B = 1 \mid Z = 1)$ in Figure 1).

If we are given multiple pieces of virtual evidence, due to its commutative property, we can use our procedure in Theorem 13, with some slight changes, to incorporate the pieces of virtual evidence in any order, which we give a sketch here. After having already incorporated the first piece of virtual evidence $Z^1 = 1$, if we want to incorporate the second piece of virtual evidence $Z^2 = 1$, steps 1 to 3 should be modified such that, instead of the default case $\Lambda^0(\top)$ where no evidence has been set, we now set the values of the indicators as $\Lambda^1(\top)$, i.e., $\lambda_{X=x} \leftarrow L^1(X = x)$ as computed by step 5 for virtual evidence $Z^1 = 1$ (since we have incorporated virtual evidence $Z^1 = 1$). The procedure continues so on for the remaining pieces of virtual evidence, where we must take into account of all previous pieces of virtual evidence that have been incorporated. Note that if we have multiple pieces of virtual evidence on the same variable X, then the indicator should be updated by multiplying into its original value, i.e., $\lambda_{X=x} \leftarrow \lambda_{X=x} L^1(X = x)$ for example.

Finally, we note that the method of incorporating virtual evidence into a Bayesian network in Corollary 6 is equivalent to Theorem 13. This is because if we add a new variable Z as a child of X, we are in effect modifying the arithmetic circuit by replacing the original indicator node $\lambda_{X=x}$ with a multiplication node of two children, where one child is the new indicator node $\lambda_{X=x}$, and the other child is a product of terms, that after conditioning on $Z = 1$ and evaluating accordingly, can be shown to compute the same as Equation 4.

6. Conclusion

In this paper, we showed how to incorporate uncertain evidence in the form of virtual evidence to a decomposable and smooth arithmetic circuit representing a probability distribution using a procedure which sets the indicators to non-negative values (other than 0 or 1) and then runs tractable operations of upward and downward passes, which takes time linear in the size of the arithmetic circuit.

In the future, we will look at incorporating uncertain evidence into arithmetic circuits which satisfy more or fewer properties than decomposability and smoothness. There are two reasons. The first reason is that arithmetic circuits of smaller sizes may not satisfy the properties of decomposability and smoothness, and we would like to find a procedure that allows us to incorporate uncertain evidence effectively. The second reason is that if we want to compute queries other than probabilities after incorporating uncertain evidence, the arithmetic circuit may need to satisfy further properties. For example, to find the most probable explanation (MPE) (Chan and Darwiche, 2006), i.e., the full instantiation with the highest probability given some evidence (including virtual evidence), the arithmetic circuit also needs to satisfy the property of determinism (Choi and Darwiche, 2017). We would also like to look at the problem of finding the maximum a posteriori probability (MAP) (Park and Darwiche, 2004), which are proved to NP^{PP}-complete (as opposed to PP-complete for MAP and NP-complete for marginal probability), after incorporating uncertain evidence.

Acknowledgments

This research is partly supported by JSPS KAKENHI(S) Grant Number 15H05711.

Appendix: proof sketches

Before we proceed with the proofs, we define a *complete sub-circuit* of an arithmetic circuit that is both decomposable and smooth (Chan and Darwiche, 2006; Choi and Darwiche, 2017).

Definition 14 *A complete sub-circuit of an arithmetic circuit is obtained by traversing the circuit top-down, while choosing one child of each visited +-node and all children of each visited *-node.*

Similar to an arithmetic circuit, we can consider the output of a complete sub-circuit by traversing the sub-circuit bottom-up and compute the value of the root node. It can be shown that the output of an arithmetic circuit is the sum of the outputs of all complete sub-circuits (Choi and Darwiche, 2017). Moreover, given $X = x$, we say a complete sub-circuit is x-consistent if $\lambda_{X=x}$ is a leaf node in the complete sub-circuit, and not x-consistent otherwise.

Corollary 15 *If the output of an arithmetic circuit is $f(\mathbf{y})$, then the sum of the outputs of all complete sub-circuits that are x-consistent is $f(\mathbf{y}, x)$ and the sum of the outputs of all complete sub-circuits that are not x-consistent is $f(\mathbf{y}) - f(\mathbf{y}, x)$ (Choi and Darwiche, 2017).*

Proof of Theorem 12

For each node n, there are two registers:

- $vr(n)$ to store the partial values of the arithmetic circuit during the upward pass
- $dr(n)$ to store the partial derivatives of the arithmetic circuit during the downward pass

We make two observations. First, when computing the dr value for each node n during the downward pass, its dr value depends on the dr values of its parents, and recursively its ancestors only. Therefore, as our proof only considers the dr value of the indicator node $\lambda_{X=x}$, we can ignore the parts of the algorithm which compute the dr values of non-ancestors of the indicator node $\lambda_{X=x}$.

Second, we can modify the arithmetic circuit without having any impact on the final vr value of the root (i.e., the output of the arithmetic circuit) or the final dr value of the leaf nodes (i.e., the partial derivatives with respect to the leaf node). First, if two nodes are parent and child and are of the same type (i.e., both +-nodes or both *-nodes), we can combine them together (meaning that each path in the arithmetic circuit must now alternate between +-nodes and *-nodes). Second, if an internal node has multiple parents, we can split this internal node into multiple nodes such that each internal node has only one parent (meaning that only leaf nodes are allowed to have multiple parents).

We now set the current nodes, denoted as v, and the sum of the product of dr and vr of all current nodes, denoted as $p = \sum_v dr(v)vr(v)$. We start with the root node as the current nodes, which is initialized as $vr(root) = f(\mathbf{y})$ and $dr(root) = 1$, meaning $p = dr(root)vr(root) = f(\mathbf{y})$. Since our arithmetic circuit alternates between +-nodes and

*-nodes, all current nodes are either all +-nodes or *-nodes if we go down one level, before we reach the indicator node $\lambda_{X=x}$.

If the current nodes are *-nodes, due to decomposability, for each current node v at most one of its children is an ancestor of $\lambda_{X=x}$, which we denote this child as c, and all other children that are not ancestors of $\lambda_{X=x}$ are denoted as c'. By the algorithm of the downward pass, $dr(c) \leftarrow dr(c) + dr(v)vr'$, where $vr' = \prod_{c'} vr(c')$. By the algorithm of the upward pass we have finished, we know that $vr(n) = vr(c)vr'$. Since $dr(c)$ is initialized as 0 and c has one parent only, this means $dr(c) = dr(v)vr'$ and $dr(c)vr(c) = dr(v)vr(c)vr' = dr(v)vr(v)$. We then choose c as the new current nodes to replace v, and this means the new p value is unchanged from before.

If the current nodes are +-nodes, due to smoothness, for each current node v at least one of its children is an ancestor of $\lambda_{X=x}$, which we denote these children as c, and all other children that are not ancestors of $\lambda_{X=x}$ are denoted as c'. By the algorithm of the downward pass, $dr(c) \leftarrow dr(c) + dr(v)$. Since $dr(c)$ is initialized as 0 and c has one parent only, this means $dr(c) = dr(v)$. By the algorithm of the upward pass we have finished, we know that $vr(v) = \sum_c vr(c) + \sum_{c'} vr(c')$. We then choose c as the new current nodes to replace v, and this means the new p value is decreased from before. In fact, we have removed $\sum_{c'} vr(c')$, where we can consider as the sum of the outputs of a subset of complete sub-circuits that are not x-consistent.

Throughout the downward pass, we visit the arithmetic circuit top-down, alternating between *-nodes (which keeps p unchanged) and +-nodes (which decreases p by effectively removing the sum of the outputs of a subset of complete sub-circuits that are not x-consistent). When we reach the indicator node $\lambda_{X=x}$, all outputs of complete sub-circuits that are not x-consistent are removed from p, keeping only the outputs of complete sub-circuits that are x-consistent, which sums to $f(\mathbf{y}, x)$. Therefore, when we reach the indicator node $\lambda_{X=x}$, we have $p = f(\mathbf{y}, x)$. Since $p = dr(\lambda_{X=x})vr(\lambda_{X=x})$ and $vr(\lambda_{X=x}) = \lambda_{X=x}$, we have $\partial \mathcal{AC}(\Lambda(\mathbf{y}))/\partial\lambda_{X=x} = dr(\lambda_{X=x}) = p/vr(\lambda_{X=x}) = f(\mathbf{y}, x)/\lambda_{X=x}$.

Proof of Theorem 13

In the arithmetic circuit, the sum of the outputs of complete sub-circuits that are $\mathbf{x}$-consistent is $Pr(\mathbf{x})$. Due to decomposability and smoothness, the output of each of these sub-circuits that are $\mathbf{x}$-consistent is in effect a product of terms, where one and only one of the terms is the indicator $\lambda_{X=x}$ where $x \sim \mathbf{x}$. Since the value of $\lambda_{X=x}$ is changed from 1 to $L^1(X = x)$ (from Equation 4), this in effect computes the product $L^1(X = x)Pr(\mathbf{x})$, which is equivalent to the incorporating virtual evidence in Equation 3.

For $\Pr(\mathbf{y})$, it is the sum of $Pr(\mathbf{x})$ for all instantiations $\mathbf{x}$ that are consistent with $\mathbf{y}$, which for each instantiation have probability multiplied by the corresponding $L^1(X = x)Pr(\mathbf{x})$ after incorporating virtual evidence. It can also be computed by the arithmetic circuit, $\mathcal{AC}(\Lambda(\mathbf{y}))$, which is the sum of the complete sub-circuits that are $\mathbf{y}$-consistent, where each of them has the value of $\lambda_{X=x}$ changed from 1 to $L^1(X = x)$. Due to this one-to-one correspondence, the procedure of Theorem 13 computes $\mathcal{AC}(\Lambda^1(\mathbf{y})) = Pr(\mathbf{y} \mid Z = 1)$ during the upward pass (Step 6), and $\partial \mathcal{AC}(\Lambda^1(\mathbf{y}))/\partial\lambda_{X_i=x_i} = Pr(\mathbf{y}, X_i = x_i \mid Z = 1)/\lambda_{X_i=x_i}$ for all values x_i of all variables $X_i \notin \mathbf{Y}$ during the downward pass (Step 7).

References

Ali Ben Mrad, Vronique Delcroix, Sylvain Piechowiak, Philip Leicester, and Mohamed Abid. An explication of uncertain evidence in Bayesian networks: Likelihood evidence and probabilistic evidence. *Applied Intelligence*, 43(4):802–824, 2015.

Hei Chan and Adnan Darwiche. On the revision of probabilistic beliefs using uncertain evidence. *Artificial Intelligence*, 163:67–90, 2005.

Hei Chan and Adnan Darwiche. On the robustness of most probable explanations. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 63–71, 2006.

Arthur Choi and Adnan Darwiche. On relaxing determinism in arithmetic circuits. In *Proceedings of the the 34th International Conference on Machine Learning (ICML)*, 2017.

Adnan Darwiche. Decomposable negation normal form. *Journal of the ACM*, 48(4):608–647, 2001.

Adnan Darwiche. A differential approach to inference in Bayesian networks. *Journal of the ACM*, 50(3):280–305, 2003.

Adnan Darwiche. *Modeling and Reasoning with Bayesian Networks*. Cambridge University Press, 2009.

Richard Jeffrey. *The Logic of Decision*. University of Chicago Press, 2nd edition, 1990.

Doga Kisa, Guy Van den Broeck, Arthur Choi, and Adnan Darwiche. Probabilistic sentential decision diagrams. In *Proceedings of the 14th International Conference on Principles of Knowledge Representation and Reasoning (KR)*, 2014.

Daniel Lowd and Pedro M. Domingos. Learning arithmetic circuits. In *Proceedings of the 24th Conference in Uncertainty in Artificial Intelligence (UAI)*, pages 383–392, 2008.

James D. Park and Adnan Darwiche. Complexity results and approximation strategies for map explanations. *Journal of Artificial Intelligence Research*, 21:101–133, 2004.

Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

Hoifung Poon and Pedro M. Domingos. Sum-product networks: A new deep architecture. In *Proceedings of the 27th Conference in Uncertainty in Artificial Intelligence (UAI)*, pages 373–346, 2011.

Carl Wagner. Probability kinematics and commutativity. *Philosophy of Science*, 69:266–278, 2002.