

On The Role of Contextual Weak Independence in Probabilistic Inference

C.J. Butz and M.J. Sanscartier

Department of Computer Science, University of Regina
Regina, SK, S4S 0A2, Canada

Abstract. Previous experimental results have clearly demonstrated the effectiveness of utilizing *context-specific independence* (CSI) in probabilistic inference. However, CSI is a special case of a more general independence called *contextual weak independence* (CWI). In this paper, we show how CWI can be utilized for more efficient probabilistic inference. These results are quite significant as they suggest that CWI may play an important role in probabilistic inference.

1 Introduction

In practice, probabilistic inference would not be feasible without making independency assumptions. Directly specifying a joint probability distribution is not always possible as one would have to specify 2^n entries for a distribution over n binary variables. However, *Bayesian networks* [2, 4] have become a basis for designing probabilistic expert systems as the *conditional independence* (CI) assumptions encoded in a Bayesian network allow for a joint distribution to be *indirectly* specified as a product of *conditional probability tables* (CPTs). More importantly, perhaps, this factorization can lead to computationally feasible inference in some applications. Nevertheless, this approach to probabilistic inference is rather limited since it is based on a very strict type of independence, the probabilistic conditional independence.

It is well-known that the notion of conditional independence is too restrictive to capture independencies that only hold in certain contexts. This kind of contextual independency was formalized as *context-specific independence* (CSI) by Boutilier et al. [1]. The important point is that Zhang and Poole [7] have empirically demonstrated that CSI can significantly speed up inference. At the same time, Wong and Butz [6] emphasized that CSI is a special case of a more general contextual independency called *contextual weak independence* (CWI).

In this paper, we show that CWI may be more useful than CSI in probabilistic inference. While the notion of CI can factorize a joint distribution as a product of CPTs, the notion of CSI can refine the CPTs themselves. Since CSI is a special case of CWI, the notion of CWI can further refine the CPTs. We explicitly demonstrate in Section 4 that this refinement can *reduce* the number of multiplications and additions needed for probabilistic inference. Finally, it is worth mentioning that although this paper focuses on inference using CWI, we take advantage of the *union product* operator developed by Zhang and Poole [7].

This paper is organized as follows. In Section 2, we briefly review probabilistic inference in Bayesian networks. In Section 3, we illustrate the usefulness of CSI in inference. In Section 4, we show how more efficient probabilistic inference can be achieved in a CWI approach using independencies that would go unnoticed in a CSI approach. The conclusion is presented in Section 5.

2 Bayesian Networks

Consider a finite set $U = \{A_1, A_2, \dots, A_n\}$ of discrete random variables, where each variable $A \in U$ takes on values from a finite domain V_A . We may use capital letters, such as A, B, C , for variable names and lowercase letters a, b, c to denote specific values taken by those variables. Sets of variables will be denoted by capital letters such as X, Y, Z , and assignments of values to the variables in these sets (called configurations or tuples) will be denoted by lowercase letters x, y, z . We use V_X in the obvious way. We shall also use the short notation $p(a)$ for the probabilities $p(A = a)$, $a \in V_A$, and $p(z)$ for the set of variables $Z = \{A, B\} = AB$ meaning $p(Z = z) = p(A = a, B = b) = p(a, b)$, where $a \in V_A, b \in V_B$.

Let p be a *joint probability distribution* (jpd) [2] over the variables in U and X, Y, Z be subsets of U . We say Y and Z are *conditionally independent* given X , if given any $x \in V_X, y \in V_Y$, then for all $z \in V_Z$,

$$p(y \mid x, z) = p(y \mid x), \quad \text{whenever } p(x, z) > 0. \quad (1)$$

For convenience we write Eq. (1) as $p(Y \mid X, Z) = p(Y \mid X)$.

Based on the *conditional independence* (CI) assumptions encoded in the Bayesian network in Fig. 1, the jpd $p(A, B, C, D, E)$ can be factorized as

$$p(A, B, C, D, E) = p(A) \cdot p(B) \cdot p(C|A) \cdot p(D|A, B) \cdot p(E|A, C, D). \quad (2)$$

Using the CPTs $p(D|A, B)$ and $p(E|A, C, D)$ shown in Fig. 2, we conclude this section with an example of probabilistic inference.

The distribution $p(A, B, C, E)$ can be computed from Eq. (2) as

$$\begin{aligned} p(A, B, C, E) &= \sum_D p(A, B, C, D, E) \\ &= \sum_D p(A) \cdot p(B) \cdot p(C|A) \cdot p(D|A, B) \cdot p(E|A, C, D) \\ &= p(A) \cdot p(B) \cdot p(C|A) \cdot \sum_D p(D|A, B) \cdot p(E|A, C, D). \end{aligned} \quad (3)$$

Computing the product $p(D|A, B) \cdot p(E|A, C, D)$ of the two distributions in Fig. 2 requires 32 multiplications. Marginalizing out variable D from this product requires 16 additions. The resulting distribution can be multiplied with $p(A) \cdot p(B) \cdot p(C|A)$ to obtain our desired distribution $p(A, B, C, E)$.

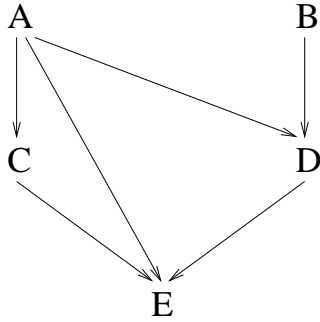


Fig. 1. A Bayesian network.

A	B	D	$p(D A, B)$	A	C	D	E	$p(E A, C, D)$
0	0	0	0.3	0	0	0	0	0.1
0	0	1	0.7	0	0	0	1	0.9
0	1	0	0.3	0	0	1	0	0.1
0	1	1	0.7	0	0	1	1	0.9
1	0	0	0.6	0	1	0	0	0.8
1	0	1	0.4	0	1	0	1	0.2
1	1	0	0.8	0	1	1	0	0.8
1	1	1	0.2	0	1	1	1	0.2
				1	0	0	0	0.6
				1	0	0	1	0.4
				1	0	1	0	0.3
				1	0	1	1	0.7
				1	1	0	0	0.6
				1	1	0	1	0.4
				1	1	1	0	0.3
				1	1	1	1	0.7

Fig. 2. The CPTs $p(D|A, B)$ and $p(E|A, C, D)$ in Eq. (2).

3 Inference with Context-Specific Independence

The Bayesian network factorization of $p(A, B, C, D, E)$ in Eq. (2) only reflects conditional independencies $p(y|x, z) = p(y|x)$ which hold for *all* $x \in V_X$. In some situations, however, the conditional independence may only hold for certain *specific* values in V_X .

Consider again the CPT $p(D|A, B)$ redrawn in Fig. 3 (i). Although variables D and B are *not* conditionally independent given A , it can be seen in Fig. 3 (ii,iii) that D and B are independent in context $A = 0$, that is,

$$p(D = d|A = 0, B = b) = p(D = d|A = 0).$$

Similarly, for the CPT $p(E|A, C, D)$ redrawn in Fig. 4 (i), it can be seen in Fig. 4 (ii,iii) that variables E and D are independent given C in context $A = 0$, while variables E and C are independent given D in context $A = 1$, i.e.,

$$p(E = e|A = 0, C = c, D = d) = p(E = e|A = 0, C = c)$$

and

$$p(E = e|A = 1, C = c, D = d) = p(E = e|A = 1, D = d).$$

<table style="border-collapse: collapse; width: 100%;"> <thead> <tr><th>A</th><th>B</th><th>D</th><th>$p(D A, B)$</th></tr> </thead> <tbody> <tr><td>0</td><td>0</td><td>0</td><td>0.3</td></tr> <tr><td>0</td><td>0</td><td>1</td><td>0.7</td></tr> <tr><td>0</td><td>1</td><td>0</td><td>0.3</td></tr> <tr><td>0</td><td>1</td><td>1</td><td>0.7</td></tr> <tr><td>1</td><td>0</td><td>0</td><td>0.6</td></tr> <tr><td>1</td><td>0</td><td>1</td><td>0.4</td></tr> <tr><td>1</td><td>1</td><td>0</td><td>0.8</td></tr> <tr><td>1</td><td>1</td><td>1</td><td>0.2</td></tr> </tbody> </table>	A	B	D	$p(D A, B)$	0	0	0	0.3	0	0	1	0.7	0	1	0	0.3	0	1	1	0.7	1	0	0	0.6	1	0	1	0.4	1	1	0	0.8	1	1	1	0.2	<table style="border-collapse: collapse; width: 100%;"> <thead> <tr><th>A</th><th>B</th><th>D</th><th>$p(D A = 0, B)$</th></tr> </thead> <tbody> <tr><td>0</td><td>0</td><td>0</td><td>0.3</td></tr> <tr><td>0</td><td>0</td><td>1</td><td>0.7</td></tr> <tr><td>0</td><td>1</td><td>0</td><td>0.3</td></tr> <tr><td>0</td><td>1</td><td>1</td><td>0.7</td></tr> </tbody> </table>	A	B	D	$p(D A = 0, B)$	0	0	0	0.3	0	0	1	0.7	0	1	0	0.3	0	1	1	0.7	→	<table style="border-collapse: collapse; width: 100%;"> <thead> <tr><th>A</th><th>D</th><th>$p(D A = 0)$</th></tr> </thead> <tbody> <tr><td>0</td><td>0</td><td>0.3</td></tr> <tr><td>0</td><td>1</td><td>0.7</td></tr> </tbody> </table>	A	D	$p(D A = 0)$	0	0	0.3	0	1	0.7
A	B	D	$p(D A, B)$																																																																	
0	0	0	0.3																																																																	
0	0	1	0.7																																																																	
0	1	0	0.3																																																																	
0	1	1	0.7																																																																	
1	0	0	0.6																																																																	
1	0	1	0.4																																																																	
1	1	0	0.8																																																																	
1	1	1	0.2																																																																	
A	B	D	$p(D A = 0, B)$																																																																	
0	0	0	0.3																																																																	
0	0	1	0.7																																																																	
0	1	0	0.3																																																																	
0	1	1	0.7																																																																	
A	D	$p(D A = 0)$																																																																		
0	0	0.3																																																																		
0	1	0.7																																																																		
(i)	(ii)		(iii)																																																																	

Fig. 3. Variables D and B are conditionally independent in context $A = 0$.

This kind of contextual independency was formalized as *context-specific independence* (CSI) by Boutilier et al. [1] as follows. Let X, Y, Z, C be pairwise disjoint subsets of U and $c \in V_C$. We say Y and Z are *conditionally independent* given X in *context* $C = c$, if

$$p(y | x, z, c) = p(y | x, c), \quad \text{whenever } p(x, z, c) > 0.$$

In order to utilize the above three context-specific independencies for more efficient probabilistic inference, Zhang and Poole [7] generalized the standard product operator \cdot as the *union product* operator \odot . The *union product* $p(Y, X) \odot q(X, Z)$ of functions $p(Y, X)$ and $q(X, Z)$ is the function on YXZ defined as

$$p(y, x) \odot q(x, z) = \begin{cases} p(y, x) \cdot q(x, z) & \text{if both } p(y, x) \text{ and } q(x, z) \text{ are defined} \\ p(y, x) & \text{if } p(y, x) \text{ is defined and } q(x, z) \text{ is undefined} \\ q(x, z) & \text{if } p(y, x) \text{ is undefined and } q(x, z) \text{ is defined} \\ \text{undefined} & \text{if both } p(y, x) \text{ and } q(x, z) \text{ are undefined.} \end{cases}$$

Note that \odot is commutative and associative [7]. (It should be mentioned that Zhang and Poole [7] also pointed out that the notion of CSI can be applied in the problem of constructing a Bayesian network [5].)

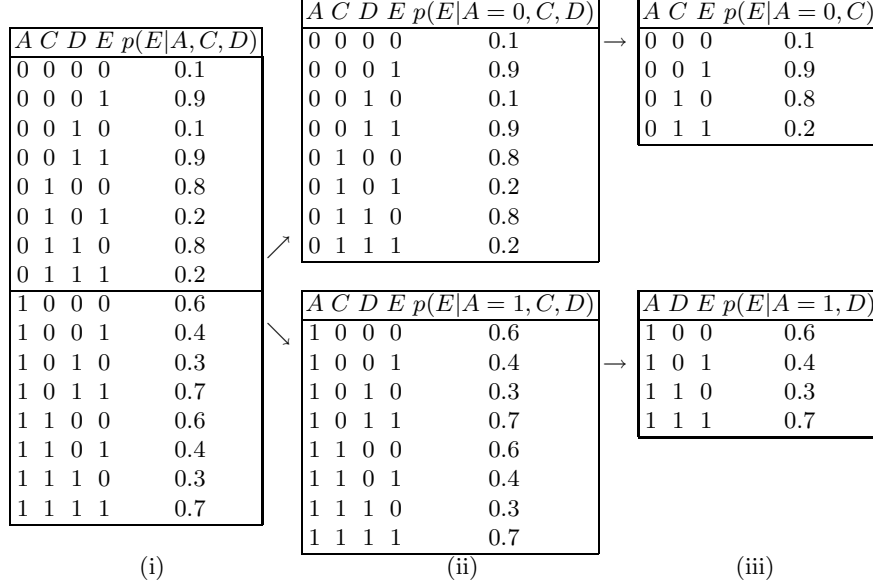


Fig. 4. Variables E and D are conditionally independent given C in context $A = 0$, while E and C are conditionally independent given D in context $A = 1$.

The union product operator allows for a single CPT to be horizontally partitioned into more than one CPT, which, in turn, exposes the contextual independencies. Returning to the factorization in Eq. (2), the CPT $p(D|A, B)$ can be rewritten as

$$\begin{aligned}
 p(D|A, B) &= p(D|A=0, B) \odot p(D|A=1, B) \\
 &= p(D|A=0) \odot p(D|A=1, B)
 \end{aligned} \tag{4}$$

while $p(E|A, C, D)$ is equivalently stated as

$$\begin{aligned}
 p(E|A, C, D) &= p(E|A=0, C, D) \odot p(E|A=1, C, D) \\
 &= p(E|A=0, C) \odot p(E|A=1, D).
 \end{aligned} \tag{5}$$

By substituting Eqs. (4) and (5) into Eq. (2), the factorization of the jpd $p(A, B, C, D, E)$ using CSI is

$$\begin{aligned}
 p(A, B, C, D, E) &= p(A) \cdot p(B) \cdot p(C|A) \odot p(D|A=0) \odot p(D|A=1, B) \\
 &\quad \odot p(E|A=0, C) \odot p(E|A=1, D).
 \end{aligned} \tag{6}$$

The use of CSI leads to more efficient probabilistic inference.

Computing $p(A, B, C, E)$ from Eq. (6) involves

$$\begin{aligned}
p(A, B, C, E) &= \sum_D p(A) \cdot p(B) \cdot p(C|A) \odot p(D|A=0) \odot p(D|A=1, B) \\
&\quad \odot p(E|A=0, C) \odot p(E|A=1, D) \\
&= p(A) \cdot p(B) \cdot p(C|A) \odot p(E|A=0, C) \odot \sum_D p(D|A=0) \\
&\quad \odot p(D|A=1, B) \odot p(E|A=1, D). \tag{7}
\end{aligned}$$

Computing the union product $p(D|A=0) \odot p(D|A=1, B) \odot p(E|A=1, D)$ requires 8 multiplications. Next, 8 additions are required to marginalize out variable D . Eight more multiplications are required to compute the union product of the resulting distribution with $p(E|A=0, C)$. The resulting distribution can be multiplied with $p(A) \cdot p(B) \cdot p(C|A)$ to give $p(A, B, C, E)$.

The important point in this section is that computing $p(A, B, C, E)$ from the CSI factorization in Eq. (7) required 16 fewer multiplications and 8 fewer additions compared to the respective number of computations needed to compute $p(A, B, C, E)$ from the CI factorization in Eq. (3).

4 Inference with Contextual Weak Independence

Since CSI is a special case of *contextual weak independence* (CWI) [6], any computational savings achieved in a CSI approach will also be achieved in a CWI approach. In addition, we show in this section that more efficient probabilistic inference can be achieved in a CWI approach using independencies that would go unnoticed in a CSI approach.

Consider another jpd $p'(A, B, C, D, E)$ which also satisfies the conditional independencies encoded in the Bayesian network in Fig. 1,

$$p'(A, B, C, D, E) = p'(A) \cdot p'(B) \cdot p'(C|A) \cdot p'(D|A, B) \cdot p'(E|A, C, D). \tag{8}$$

The two CPTs $p'(D|A, B)$ and $p'(E|A, C, D)$ are shown in Fig. 5 (i) and Fig. 6 (i), respectively.

In the CPT $p'(D|A, B)$, there are *no* context-specific independencies holding in $p'(D|A=0, B)$ nor in $p'(D|A=1, B)$. Similarly, in the CPT $p'(E|A, C, D)$, there are *no* context-specific independencies holding in $p'(E|A=0, C, D)$ nor in $p'(E|A=1, C, D)$. This means that no refinement of the BN factorization in Eq. (8) is possible in a CSI approach. Thereby, computing $p'(A, B, C, E)$ from Eq. (8) in a CSI approach involves

$$\begin{aligned}
p'(A, B, C, E) &= \sum_D p'(A) \cdot p'(B) \cdot p'(C|A) \cdot p'(D|A, B) \cdot p'(E|A, C, D) \\
&= p'(A) \cdot p'(B) \cdot p'(C|A) \cdot \sum_D p'(D|A, B) \cdot p'(E|A, C, D). \tag{9}
\end{aligned}$$

Computing $\sum_D p'(D|A, B) \cdot p'(E|A, C, D)$ requires 64 multiplications and 32 additions.

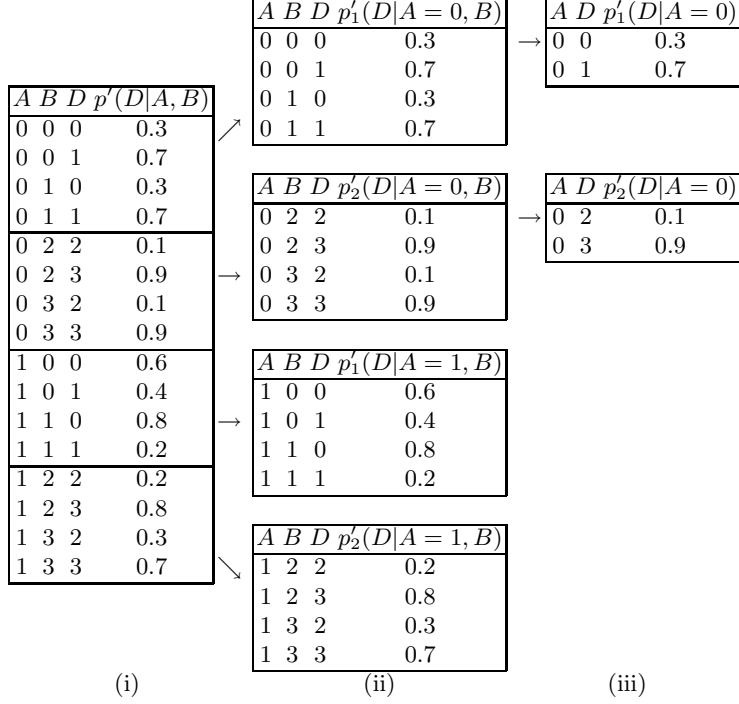


Fig. 5. Variables D and B are *weakly* independent in context $A = 0$.

Unlike the definition of CSI [1], the definitions of CWI (given below) and the union product operator \odot do *not* require a CPT to be horizontally partitioned as a dichotomy (see Fig. 3 and 4). On the contrary, by the definition of \odot , the CPT $p'(D|A, B)$ in Eq. (8) can be written as

$$\begin{aligned}
 & p'(D|A, B) \\
 = & p'_1(D|A = 0, B) \odot p'_2(D|A = 0, B) \odot p'_1(D|A = 1, B) \odot p'_2(D|A = 1, B), \quad (10)
 \end{aligned}$$

as illustrated in Fig. 5 (i,ii). Variables D and B are conditionally independent in context $A = 0$ in both $p'_1(D|A = 0, B)$ and $p'_2(D|A = 0, B)$, as depicted in Fig. 5 (iii). Thus, Eq. (10) can be refined as

$$\begin{aligned}
 & p'(D|A, B) \\
 = & p'_1(D|A = 0) \odot p'_2(D|A = 0) \odot p'_1(D|A = 1, B) \odot p'_2(D|A = 1, B). \quad (11)
 \end{aligned}$$

Similarly, the CPT $p'(E|A, C, D)$ can also be factorized as

$$\begin{aligned}
 p'(E|A, C, D) = & p'_1(E|A = 0, C, D) \odot p'_2(E|A = 0, C, D) \\
 & \odot p'_1(E|A = 1, C, D) \odot p'_2(E|A = 1, C, D), \quad (12)
 \end{aligned}$$

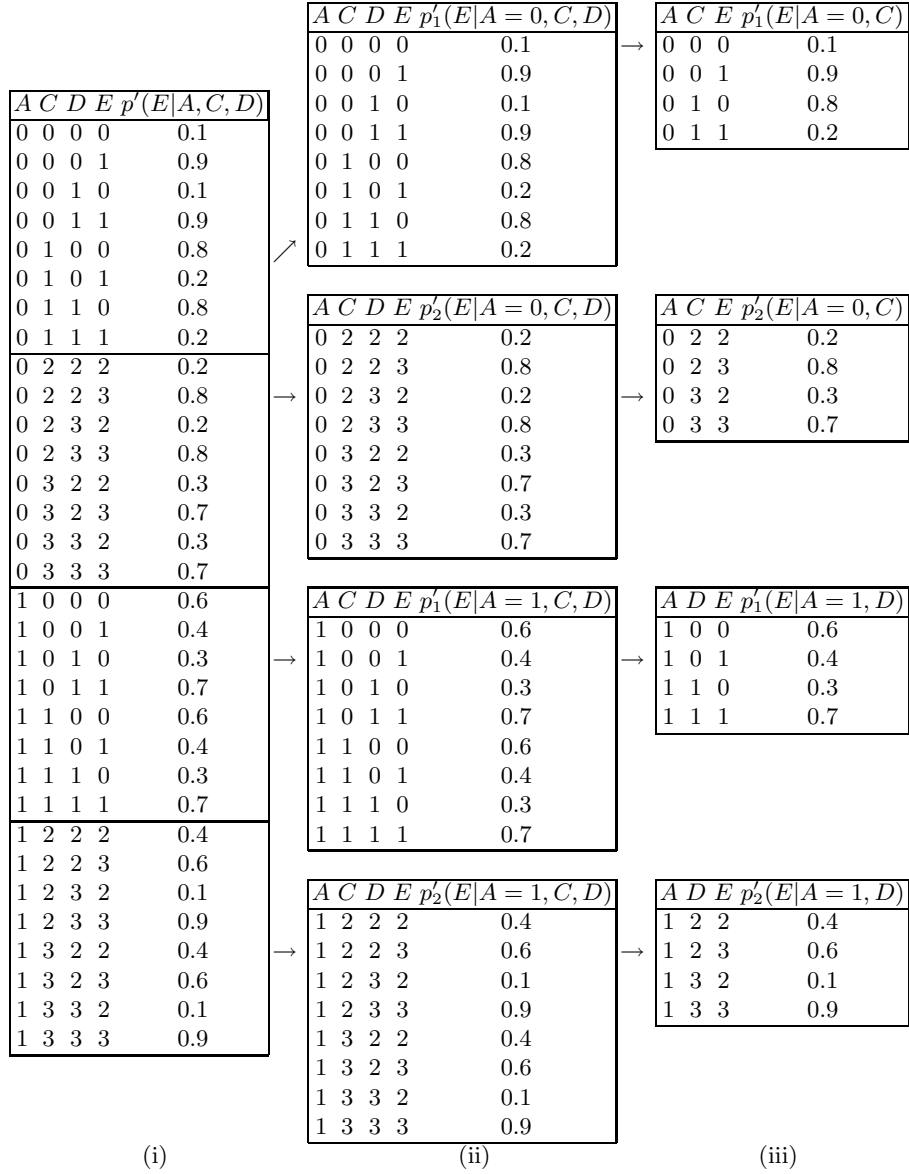


Fig. 6. Variables E and D are *weakly* independent given C in context $A = 0$, while E and C are *weakly* independent given D in context $A = 1$.

as illustrated in Fig. 6 (i,ii). Variables E and D are conditionally independent given C in context $A = 0$ in both $p'_1(E|A = 0, C, D)$ and $p'_2(E|A = 0, C, D)$, as depicted in Fig. 6 (iii). In addition, E and C are conditionally independent given D in context $A = 1$ in both $p'_1(E|A = 1, C, D)$ and $p'_2(E|A = 1, C, D)$, as shown in Fig. 6 (iii). These independencies can be used to refine Eq. (12) as

$$\begin{aligned} & p'(E|A, C, D) \\ = & p'_1(E|A = 0, C) \odot p'_2(E|A = 0, C) \odot p'_1(E|A = 1, D) \odot p'_2(E|A = 1, D). \end{aligned} \quad (13)$$

This type of contextual independency was formalized as *contextual weak independence* (CWI) by Wong and Butz [6] as follows. Let X, Y, Z, C be pairwise disjoint subsets of U and $c \in V_C$. We say Y and Z are *weakly independent* given X in context $C = c$, if both of the following two conditions are satisfied: (i) there exists a *maximal disjoint compatibility class* [3] $\pi = \{t_i, \dots, t_j\}$ in the relation $\theta(X, Y, C = c) \circ \theta(X, Z, C = c)$, and (ii) given any $x \in V_X^\pi$, $y \in V_Y^\pi$, then for all $z \in V_Z^\pi$,

$$p(y | x, z, c) = p(y | x, c), \text{ whenever } p(x, z, c) > 0,$$

where $\theta(W)$ denotes the equivalence relation induced by the set W of variables, \circ denotes the composition operator, and V_W^π denotes the set of values for W appearing in π .

Unlike the notion of CSI, the notion of CWI can *refine* the Bayesian network factorization of $p'(A, B, C, D, E)$ in Eq. (8). By substituting Eqs. (11) and (13) into Eq. (8), the factorization of $p'(A, B, C, D, E)$ in a CWI approach is

$$\begin{aligned} & p'(A, B, C, D, E) \\ = & p'(A) \cdot p'(B) \cdot p'(C|A) \odot p'_1(D|A = 0) \odot p'_2(D|A = 0) \odot p'_1(D|A = 1, B) \\ & \odot p'_2(D|A = 1, B) \odot p'_1(E|A = 0, C) \odot p'_2(E|A = 0, C) \\ & \odot p'_1(E|A = 1, D) \odot p'_2(E|A = 1, D). \end{aligned} \quad (14)$$

Computing $p'(A, B, C, E)$ from Eq. (14) involves

$$\begin{aligned} & p'(A, B, C, E) \\ = & \sum_D p'(A) \cdot p'(B) \cdot p'(C|A) \odot p'_1(D|A = 0) \odot p'_2(D|A = 0) \odot p'_1(D|A = 1, B) \\ & \odot p'_2(D|A = 1, B) \odot p'_1(E|A = 0, C) \odot p'_2(E|A = 0, C) \\ & \odot p'_1(E|A = 1, D) \odot p'_2(E|A = 1, D) \\ = & p'(A) \cdot p'(B) \cdot p'(C|A) \odot p'_1(E|A = 0, C) \odot p'_2(E|A = 0, C) \\ & \odot \sum_D p'_1(D|A = 0) \odot p'_2(D|A = 0) \odot p'_1(D|A = 1, B) \odot p'_2(D|A = 1, B) \\ & \odot p'_1(E|A = 1, D) \odot p'_2(E|A = 1, D). \end{aligned} \quad (15)$$

In this case, only 32 multiplications and 16 additions are required to compute the distribution to be multiplied with $p'(A) \cdot p'(B) \cdot p'(C|A)$, as opposed to the needed 64 multiplications and 32 additions in the CSI factorization in Eq. (9).

The main point in this section is that computing $p'(A, B, C, E)$ from the CWI factorization in Eq. (15) required 32 fewer multiplications and 16 fewer additions compared to the respective number of computations needed to compute $p'(A, B, C, E)$ in the CSI factorization in Eq. (9).

5 Conclusion

Recently, it has been empirically demonstrated that CSI can lead to more efficient probabilistic inference than can be obtained using CI alone [7]. At the same time, it has been shown in [6] that CSI is a *special case* of CWI. This means that any computational savings achieved in a CSI approach will also be achieved in a CWI approach. In addition, as shown in Section 4, more efficient probabilistic inference can be obtained in a CWI approach when compared to a CSI approach. We are currently conducting more thorough experiments to support the encouraging results in this paper.

References

1. Bouilrier, C., Friedman, N., Goldszmidt, M., Koller, D.: Context-specific independence in Bayesian networks, *Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence*, 115–123, 1996.
2. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann Publishers, San Francisco (1988)
3. Preparata, F. and Yeh, R.: Introduction to Discrete Structures. Addison-Wesley, Don Mills, Ontario (1973)
4. Wong, S.K.M., Butz, C.J., Wu, D.: On the implication problem for probabilistic conditional independency. *IEEE Trans. Syst. Man Cybern. SMC-A* **30**(6) (2000) 785–805
5. Wong, S.K.M., Butz, C.J.: Constructing the dependency structure of a multi-agent probabilistic network. *IEEE Trans. Knowl. Data Eng.* **13**(3) (2001) 395–415
6. Wong, S.K.M., Butz, C.J.: Contextual weak independence in Bayesian networks, *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, 670–679, 1999.
7. Zhang, N. and Poole, D.: On the role of context-specific independence in probabilistic inference, *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, 1288–1293, 1999.