Probability Bracket Notation, Multivariable Systems and Static Bayesian Networks

Dr. Xing M. Wang
Sherman Visual Lab, Sunnyvale, CA 94085, USA

Table of Contents

Abstract                   page 1
1. Introduction                       page 1
   1.1. Probability Distribution of One-Variable Systems                       page 2
   1.2. Probability Distributions of Two-Variable Systems                       page 2
   1.3. Bayes’ Chain Rule of Multivariable Systems            page 3
2. Probability Bracket Notation and Multivariable Systems                       page 4
   2.1. The Probability Basis of One-Variable Systems                      page 4
   2.2. Two-Variable Systems and the Student I-S Example            page 5
   2.3. Three-Variable Systems and the Student D-I-G Example            page 11
   2.4. Properties of Conditional Independence            page 14
   2.5. Definitions and Rules on Bayesian Networks (Informal)            page 15
3. The Student Bayesian Network and Reasoning Examples            page 15
   3.1. Local Independences of Student BN and Some Reasoning Results      page 15
   3.2. Homework Exercises: Deriving Reasoning Results for student BN      page 18
   3.3. Graphic Presentation of Student BN by Using Elvira                       page 21
Summary                       page 24
References           page 24

Abstract

Probability Bracket Notation (PBN) is applied to systems of multiple random variables for preliminary study of static Bayesian Networks (BN) and Probabilistic Graphic Models (PGM). The famous Student BN Example is explored to show the local independences and reasoning power of a BN. Software package Elvira is used to graphically display the student BN. Our investigation shows that PBN provides a consistent and convenient alternative to manipulate many expressions related to joint, marginal and conditional probability distributions in static BN.

1. Introduction

Inspired by the Dirac notation, Probability Bracket Notation (PBN) was proposed [1] for probability modeling. Now we want to apply PBN to study discrete multivariable systems in static Bayesian networks (BN) [2-6], extending our previous work on systems with mutually independent variables (see §2.3 of [1]) or induced from quantum models (see §4 of [7]).
In this section, we will briefly introduce probability distribution functions (PDF) such as joint, marginal and conditional probability distributions (JPD, MPD and CPD) [2] as well as Bayes’ chain rule for multivariable systems [3]. In next section, we will discuss the relations between miscellaneous PDF of a static BN in more details by using PBN. In the last section, we will study the local independences and reasoning power the famous Student BN Example [2-6] as our homework exercise. Finally, we will use software package Elvira [9] to graphically display and confirm our inferences on student BN Example.

For simplicity, we will not discuss topics related to dynamic Bayesian network (DBN) or Markov network [3]; and we will concentrate on discrete systems, except for a short note (see §2.2) on expressions for continuous systems in PBN [1] [7].

1.1 Probability Distribution of One-Variable Systems

A probability space \((\Omega, X, P)\) of a discrete random variable (observable) \(X\) (not necessary a real variable) is defined in PBN as follows [1]:

\[
\forall x_i \in \Omega, \quad P(x_i) \geq 0, \quad \sum_{i} P(x_i) = 1
\]  

(1.1.1)

Here the sample space \(\Omega\) (or domain) is the set of all elementary events \(x\), associated with the observable \(X\). The marginal probability distribution (MPD) can be written as [1]:

\[
P(x) = \sum_{x_i \in \Omega} P(x_i|x) = 1
\]  

(1.1.2)

The conditional probability of event \(X = x\) given \(X = x'\) can be evaluated as [1]:

\[
P(x|x') = P(X = x | X = x') = P_{X|x'}(x|x') = \frac{P(x \cap x')}{P(x')} = \frac{|| x \cap x' ||}{|| x' ||} = \delta_{x,x'}
\]  

(1.1.3)

Here we have assumed that size (count or volume) \(|| x' || \neq 0\). Then Bayes’ rule reads:

\[
P(x|x') = \frac{P(x'|x)P(x)}{P(x')} = \delta_{x',x}, \text{ for } P(x') \neq 0
\]  

(1.1.4)

Note: For a 1-variable system, there is no joint distribution function and Eq. (1.1.3-4) rather represent the orthonormality of the system basis (see §2.1).

1.2 Probability Distributions of Two-Variable Systems

Suppose \(X\) and \(Y\) are two discrete random variables in a joint probability space:

\[
(\Omega = \Omega_{X,Y} = \{\Omega_x, \Omega_y\}, \{X, Y\}, P)
\]
Then we have the following joint probability distribution (JPD) for \( \{x, y\} \in \Omega \):

\[
P(x, y) = P(x \mid y) = P(x, y) = P_{x,y}(x, y), \quad \sum_{x,y} P(x, y) = 1 \tag{1.2.1}
\]

The conditional probability distribution (CPD) of event \( X = x \) given \( Y = y \) is defined for \( P(y) \neq 0 \) as [1] [2]:

\[
P(X = x \mid Y = y) = P(x \mid y) = P_{x \mid y}(x \mid y) = \frac{P(x \cap y)}{P(y)} = \frac{P(x, y)}{P(y)} = \frac{\| x \cap y \|}{\| y \|} \tag{1.2.2}
\]

Bayes’ rule reads:

\[
P(x \mid y) = \frac{1}{P(y)} P(y \mid x) P(x), \quad \text{for } P(y) \neq 0 \tag{1.2.3}
\]

From Eq. (1.2.2) and (1.2.3) we have following two-variable chain rule:

\[
P(x, y) = P(x \mid y) P(y) = P(y \mid x) P(x) \tag{1.2.4}
\]

The two MPD in the joint probability space are given by:

\[
P(x) = \sum_y P(x, y), \quad P(y) = \sum_x P(x, y) \tag{1.2.5}
\]

### 1.3 Bayes’ Chain Rule of Multivariable Systems

Now suppose we have \( k \) discrete random variables \( X_1, X_2, \ldots, X_k \). Then what is the probability when \( X_i = x_i, \ x_i \in \Omega_i \) for \( i = 1, 2, \ldots, k \)? According to the chain rule of conditional probabilities [3], we have the following expansion of the (full) JPD:

\[
P(x_1 \cap \ldots \cap x_k) = P(x_i) P(x_2 \mid x_1) \ldots P(x_k \mid x_1 \cap \ldots \cap x_{k-1}) \tag{1.3.1}
\]

Proof (in PBN): Assuming size \( \| x_1 \cap x_2 \ldots \cap x_k \| \neq 0 \), we have

\[
P(x_1 \cap \ldots \cap x_k) = P(x_1 \cap \ldots \cap x_k \mid \Omega) = \frac{\| x_1 \cap \ldots \cap x_k \|}{\| \Omega \|}
\]

\[
= \frac{\| x_1 \| \| x_1 \cap x_2 \| \ldots \| x_1 \cap \ldots \cap x_k \|}{\| \Omega \| \| x_1 \| \ldots \| x_1 \cap \ldots \cap x_{k-1} \|}
\]

From now on, the full JPD of Eq. (1.3.1) will be expressed in following form:
\(P(x_1, \ldots, x_k) = P(x_1)P(x_2 | x_1) \cdots P(x_k | x_1, \ldots, x_{k-1})\) \hfill (1.3.2)

If all \(k\)-variables are independent to each other, as discussed in §2.3 of [1], then we have a naïve Bayesian model [2] [3]:

\[P(x_1, \ldots, x_k) = P(x_1)P(x_2) \cdots P(x_k)\] \hfill (1.3.3)

A very useful 3-variable chain rule can be derived from Eq. (1.3.2):

\[P(x_1, x_2, x_3) = P(x_1, x_2)P(x_3 | x_1, x_2) = P(x_1)P(x_2 | x_1)P(x_3 | x_1, x_2)\] \hfill (1.3.4)

From the full JPD in Eq. (1.3.4), we obtain the following intermediate JPD (or IPD):

\[P(x, y) = \sum_z P(x, y, z), \quad P(y, z) = \sum_x P(x, y, z), \quad P(x, z) = \sum_y P(x, y, z)\] \hfill (1.3.5)

Summing over one more variable, we obtain the following three MPD:

\[P(x) = \sum_{y, z} P(x, y, z), \quad P(y) = \sum_{x, z} P(x, y, z), \quad P(z) = \sum_{x, y} P(x, y, z)\] \hfill (1.3.6)

2. Probability Bracket Notation and Multivariable Systems

Let us discuss multi-variable probability space in more details by using PBN. We will use examples related to the famous Student Bayesian Network [3-5]. To avoid confusion, we will give a brief introduction to PBN, starting with single variable case.

2.1 The Probability Basis of One-Variable System

Assume that we have a system of one discrete random variable \(X\) and its probability space \((\Omega, X, P)\). Using PBN (see [1]), we have the following orthogonal and complete \(P\)-basis (probability basis):

For \(\forall x \in \Omega: \quad X | x = x | x \), \quad \(P(\Omega | x) = 1\) \hfill (2.1.1)

Orthonormality: \(P(x | x') = P(X = x | X = x') = \delta_{x', x}\) \hfill (2.1.2)

Completeness: \(\sum_{x \in \Omega} x | x \) \(P(x) = I_x\) \hfill (2.1.3)

The marginal probability distribution (MPD) is defined by:

\[P(x) = P_x(x) = P(x | \Omega) = P(X = x | \Omega), \quad x \in \Omega\] \hfill (2.1.4)

It has following property:
\[ P(x) = \begin{cases} 
0, & \text{if } x \notin \Omega \\
1, & \text{if } x \in \Omega 
\end{cases} \quad (2.1.5) \]

Its normalization can be naturally derived in PBN as follows:

\[ 1 = P(\Omega | \Omega) = P(\Omega | I_x | \Omega) = \sum_{x \in \Omega} P(\Omega | x)P(x | \Omega) = \sum_{x \in \Omega} P(x | \Omega) = \sum_{x \in \Omega} P(x) \quad (2.1.6) \]

For any subset \( H \subset \Omega \), the absolute probability can be calculated by:

\[ P(H) = P(H | \Omega) = P(H | I_x | \Omega) = \sum_{x \in \Omega} P(H | x)P(x | \Omega) = \sum_{x \in H} P(x) \quad (2.1.7) \]

The \textit{expectation value} of a well-defined real function \( F \) of observable \( X \) now can be expressed and derived as follows:

\[ E[F(X)] = P(\Omega | F(X) | \Omega) = P(\Omega | F(X)I_x | \Omega) = \sum_{x \in \Omega} P(\Omega | x)F(x)P(x | \Omega) = \sum_{x \in \Omega} P(x)F(x) \quad (2.1.8) \]

The \textit{conditional expectation value} of real function \( F(X) \) given a subset \( H \subset \Omega \) can be easily calculated:

\[ E[F(X) | H] = P(\Omega | F(X) | H) = P(\Omega | F(X)I_x | H) = \sum_{x \in \Omega} P(\Omega | x)F(x)P(x | H) = \sum_{x \in \Omega} F(x)P(x | H) = \sum_{x \in H} F(x) \frac{P(x \cap H)}{P(H)} = \sum_{x \in \Omega} \frac{1}{P(H)} \sum_{x \in H} F(x)P(x) \quad (2.1.9) \]

\subsection*{2.2 Two–Variable Systems and the Student I-S Example}

Suppose we have two discrete random variables \( X \) and \( Y \). Their sample spaces (domains) are \( \Omega_x \) and \( \Omega_y \), respectively and we also have a joint sample space \( \Omega = \{ \Omega_x, \Omega_y \} \), in which we have the following \( P \)-basis:

\[ \begin{align*}
&\text{For } \forall \{x, y\} \in \Omega: \quad X | x, y) = x | x, y, \quad Y | x, y) = y | x, y, \quad P(\Omega | x, y) = 1 \quad (2.2.1a) \\
&\text{Orthonormality: } \quad P(x, y | x', y') = \delta_{xx'} \delta_{yy'}; \quad (2.2.2b) \\
&\text{Completeness: } \quad \sum_{x, y \in \Omega} P(x, y | x, y) = I_{x,y} \quad (2.2.2c)
\end{align*} \]

The (full) \textit{joint probability distribution} (JPD) [2] is defined by:

\[ P(x, y) = P_{X,Y}(x, y) = P(x, y | \Omega) = P(X = x, Y = y | \Omega), \quad \text{for } \forall \{x, y\} \in \Omega \quad (2.2.3) \]
Here we denote:
\[ \{x, y\} \in \Omega \iff \{(x \in \Omega_x) \land (y \in \Omega_y)\} \]  
(2.2.4)

The full JPD has following property:
\[ P(x, y) \begin{cases} \geq 0, & \text{if } \{x, y\} \in \Omega \\ = 0, & \text{if } \{x, y\} \notin \Omega \end{cases} \]  
(2.2.5)

Its normalization can be derived as follows:
\[
1 = P(\Omega | \Omega) = P(\Omega | I_{X,Y} | \Omega) = \sum_{x,y \in \Omega} P(\Omega | x,y)P(x,y | \Omega)
= \sum_{x,y \in \Omega} P(x,y | \Omega) = \sum_{x,y \in \Omega} P(x,y)
\]  
(2.2.6)

The \textit{expectation value} of a well-defined real function \( F(X,Y) \) is given by:
\[
E[F(X,Y)] = P(\Omega | F(X,Y) | \Omega) = P(\Omega | F(X,Y)I_{X,Y} | \Omega)
= \sum_{x,y \in \Omega} P(\Omega | x,y)F(x,y)P(x,y | \Omega) = \sum_{x,y \in \Omega} P(x,y)F(x,y)
\]  
(2.2.7)

The conditional expectation value of real \( F(X,Y) \) given a subset \( H \subset \Omega \) can be easily calculated:
\[
E[F(X,Y) | H] \equiv P(\Omega | F(X,Y) | H) = P(\Omega | F(X,Y)I_{X,Y} | H)
= \sum_{x \in H} P(\Omega | x,y)F(x,y)P(x,y | H) = \sum_{\{x,y\} \in H} F(x,y)P(x,y | H)
\]  
(2.2.8)

From the JPD in Eq. (2.2.3), we can obtain the \textbf{MPD [2]} for \( x \) and \( y \):
\[ P(x) = \sum_{y} P(x, y | \Omega) = \sum_{y} P(x, y) \]  
(2.2.9)

\[ P(y) = \sum_{x} P(x, y | \Omega) = \sum_{x} P(x, y) \]  
(2.2.10)

In addition, we have the \textbf{CPD [2]} of \( X \) given \( Y \) and the \textbf{CPD} of \( Y \) given \( X \):
\[
P(x | y) \equiv P(X = x | Y = y) = P_{X|Y}(x | y), \quad P(y | x) = P_{Y|X}(y | x)
\]  
(2.2.11)

Please do not confuse Eq. (2.2.11) with Eq. (2.1.2), where both event \( x \) and evidence \( x' \) are the observable values of the same random variable \( X \).
**Definition 2.2.1**: Event \( X = x \) is independent of event \( Y = y \) in a distribution \( P \), denoted \( P \models x \perp y \), if \( P(x \mid y) = P(x) \) or if \( P(y) = 0 \) (§2.4.1 of [3]). 

\[
(2.2.12)
\]

**Definition 2.2.2**: variables \( X \) and \( Y \) are independent in a distribution \( P \), denoted \( P \models X \perp Y \), if \( P \models x \perp y \) for \( \forall x, y \in \Omega \). 

\[
(2.2.13)
\]

**Proposition 2.2.1**: \( P \models X \perp Y \) if and only if \( P(x, y) = P(x \wedge y) = P(x)P(y) \). Its proof can be found in §2.1.4.3 of Ref [3].

Graphically, it is represented as in Fig. 2.2.1:

**Fig. 2.2.1**: \( P \models X \perp Y \)

Now the joint distribution can be written as:

\[
P(X, Y) = P(X)P(Y)
\]

\[
(2.2.15)
\]

We see that it is the product of the distribution of each node in Fig. 2.2.1.

In many real cases, we do not know the joint distribution \( P(x, y) \). We need to derive it. Suppose we are given \( P(x \mid y) \) and \( P(y) \), then we can use Bayes’ chain rule to get it:

\[
P(x, y) = P(x \mid y)P(y)
\]

\[
(2.2.16)
\]

From it we can derive the MPD for \( X \):

\[
P(x \mid \Omega) = P(x) = \sum_y P(x, y) = \sum_y P(x \mid y)P(y) = \sum_y P(x \mid y)P(y \mid \Omega)
\]

\[
(2.2.17)
\]

Because Eq. (2.2.17) is true for \( \forall x \in \Omega \), we actually derived the unit operator for the corresponding marginal probability space of \((\Omega_y, Y, P)\):

\[
P(x \mid \Omega) = \sum_y P(x \mid y)P(y \mid \Omega) = P(x \mid \sum_y P(y \mid \Omega)) = P(x \mid I_y \mid \Omega)
\]

Or:

\[
P(x) = P_{x \mid \Omega}(x \mid \Omega) = P(x \mid I_y \mid \Omega), \quad I_y = \sum_y P(y \mid \Omega)
\]

\[
(2.2.18a)
\]

Similarly, we have the following unit operator for \( X \):
Or: \( P(y) = P(y \mid I_x \Omega) = P(y \mid I_x \mid \Omega) \), \( I_x = \sum_x | x \rangle P(x) \) \hspace{1cm} (2.2.18b)

Using unit operators, we can derive \textit{Conditional Probability (CP) Normalizations} like:

\[
1 = P(\Omega \mid x) = P(\Omega \mid I_y \mid x) = \sum_y P(\Omega \mid y)P(y \mid x) = \sum_y P(y \mid x) \\
1 = P(\Omega \mid y) = P(\Omega \mid I_x \mid y) = \sum_x P(\Omega \mid x)P(x \mid x) = \sum_x P(x \mid y) \\
\]

(2.2.19a) \hspace{1cm} (2.2.19b)

Let we use an example from the student Bayesian network [3], containing variables \( I \) (student Interagency scores) and \( S \) (student SAT scores). We assume \( S \) depends on \( I \) only. Such a relation is represented graphically as:

![Fig. 2.2.2: S is dependent on I](image)

The primary probability distribution \( P(I) \) and the CPD of \( P(S \mid I) \) are given in Fig. 3.1.2 of §3.1 (on page 17). The joint distribution can be derived by using Bayes’ chain rule:

\[
P(I, S) = P(I) P(S \mid I) \\
\]

(2.2.20)

It equals to the product of the distribution of each node in Fig. 2.2.2. Numerically, its 4 values can be calculated from tables in Fig. 3.1.2 on page 17. They are:

\[
P(i^0, s^0) = P(i^0) P(s^0 \mid i^0) = 0.7 \cdot 0.95 = 0.665 \\
P(i^0, s^1) = P(i^0) P(s^1 \mid i^0) = 0.7 \cdot 0.05 = 0.035 \\
P(i^1, s^0) = P(i^1) P(s^0 \mid i^1) = 0.3 \cdot 0.2 = 0.06 \\
P(i^1, s^1) = P(i^1) P(s^1 \mid i^1) = 0.3 \cdot 0.8 = 0.24 \\
\]

(2.2.21a) \hspace{1cm} (2.2.21b) \hspace{1cm} (2.2.21c) \hspace{1cm} (2.2.21d)

It is easy to check numerically that we have consistent MPD of \( P(I) \):

\[
P(i^0) = \sum_s P(i^0, s) = 0.665 + 0.035 = 0.7 \\
P(i^1) = \sum_s P(i^1, s) = 0.06 + 0.24 = 0.3 \\
\sum_{i^\mu} P(i^\mu) = 0.7 + 0.3 = 1.0 \\
\]

(2.2.22a) \hspace{1cm} (2.2.22b) \hspace{1cm} (2.2.22c)

This is not surprising, because:

\[
\sum_s P(i^\mu, s) = \sum_s P(i^\mu) P(s \mid i^\mu) = P(i^\mu) \sum_s P(s \mid i^\mu) = P(i^\mu) \\
\]

(2.2.23)
In the same way, we can derive the MPD of $P(S)$

$$P(s^\mu) = \sum_\rho P(i^\rho, s^\mu) = \sum_\rho P(i^\rho) P(s^\mu | i^\rho)$$  \hspace{1cm} (2.2.24)

Numerically, we have:

$$P(s^0) = \sum_{\rho=1}^2 P(i^\rho, s^0) = 0.665 + 0.06 = 0.725$$  \hspace{1cm} (2.2.25a)

$$P(s^1) = \sum_{\rho=1}^2 P(i^\rho, s^1) = 0.035 + 0.24 = 0.275$$  \hspace{1cm} (2.2.25b)

$$\sum_\mu P(s^\mu) = 0.725 + 0.275 = 1.0$$  \hspace{1cm} (2.2.25c)

Note that Eq. (2.2.24) reveals the unit operator $I_I$ for $(\Omega_I, I, P)$:

$$P(s^\mu) = \sum_\rho P(i^\rho, s^\mu) = \sum_\rho P(i^\rho) P(s^\mu | i^\rho) = \sum_\rho P(s^\mu | i^\rho) P(i^\rho | \Omega) = P(s^\mu | I_I | \Omega)$$

$\therefore P(s^\mu) = P(s^\mu | \Omega) = P_{SS}(s^\mu | \Omega) = P(s^\mu | I_I | \Omega), \quad I_I = \sum_\rho | i^\rho) P(i^\rho |$  \hspace{1cm} (2.2.26)

The CPD of $P(I | S)$ can be expressed as:

$$P(I | S) = \frac{P(I, S)}{P(S)}$$  \hspace{1cm} (2.2.27)

Numerically, we have:

$$P(i^0 | s^0) = \frac{P(i^0, s^0)}{P(s^0)} = \frac{0.665}{0.725} = 0.927$$  \hspace{1cm} (2.2.28a)

$$P(i^1 | s^0) = \frac{P(i^1, s^0)}{P(s^0)} = \frac{0.06}{0.725} = 0.083$$  \hspace{1cm} (2.2.28b)

$$P(i^0 | s^1) = \frac{P(i^0, s^1)}{P(s^1)} = \frac{0.035}{0.275} = 0.127$$  \hspace{1cm} (2.2.28c)

$$P(i^1 | s^1) = \frac{P(i^1, s^1)}{P(s^1)} = \frac{0.24}{0.275} = 0.873$$  \hspace{1cm} (2.2.28d)

Using Eq. (2.26) and (2.28), we can recover $P(I)$ as in Eq. (2.22):

$$P(i^0) = \sum_\mu P(i^0, s^\mu) = \sum_\mu P(s^\mu) P(i^0 | s^\mu) = 0.665 + 0.035 = 0.7$$

$$P(i^1) = \sum_\mu P(i^1, s^\mu) = \sum_\mu P(s^\mu) P(i^1 | s^\mu) = 0.06 + 0.24 = 0.3$$

Similarly, we can derive the unit operator $I_S$ for $(\Omega_S, S, P)$ as follows:
\[
P(i^\rho) = \sum_\mu P(i^\rho, s^\mu) = \sum_\mu P(i^\rho \mid s^\mu)P(s^\mu) = \sum_\mu P(i^\rho \mid s^\mu)P(s^\mu \mid \Omega) = P(i^\rho \mid I_s \mid \Omega) \]
\[
p(i^\rho) = P(i^\rho \mid \Omega) = P(i^\rho \mid I_s \mid \Omega) = \sum_\mu P(i^\rho \mid s^\mu)P(s^\mu \mid \Omega) = \sum_\mu P(i^\rho \mid s^\mu)P(s^\mu \mid I_s \mid \Omega), \quad I_s = \sum_\mu | s^\mu \rangle P(s^\mu) \tag{2.2.29}
\]

Note that the two unit operators \( I_i \) and \( I_s \) are derived from the expressions of cross-domain probability distributions, therefore, they can only be inserted into cross-domain probability brackets like:

\[
P(i^k) = P(i^k \mid \Omega) = P(i^k \mid I_s \mid \Omega) = \sum_\mu P(i^k \mid s^\mu)P(s^\mu \mid \Omega) = \sum_\mu P(i^k \mid s^\mu)P(s^\mu \mid I_s \mid \Omega) \tag{2.2.30a}
\]

\[
P(s^\mu) = P(s^\mu \mid \Omega) = P(s^\mu \mid I_i \mid \Omega) = \sum_\rho P(s^\mu \mid i^\rho)P(i^\rho \mid \Omega) = \sum_\rho P(s^\mu \mid i^\rho)P(i^\rho \mid I_i \mid \Omega) \tag{2.2.30b}
\]

\[
P(i^k \mid s^\nu) = P(i^k \mid I_s \mid s^\nu) = \sum_\mu P(i^k \mid s^\mu)P(s^\mu \mid s^\nu) = \sum_\mu P(i^k \mid s^\mu)\delta_{\mu\nu} = P(i^k \mid s^\nu) \tag{2.2.30c}
\]

Of course, \( I_i(I_s) \) can be used for manipulating probabilities in domain \( \Omega_i(\Omega_s) \) of the single variable \( I(S) \), just as mentioned in §2.1.

But we cannot insert the unit operator of one variable into the probability bracket of the \( P \)-basis of another variable. For example, one can verify following inequalities:

\[
0 = P(i^0 \mid i^1) \neq P(i^0 \mid I_s \mid i^1) = \sum_m P(i^0 \mid s^m)P(s^m \mid i^1) > 0
\]

\[
1 = P(i^0 \mid I_s \mid i^0) = \sum_m P(i^0 \mid s^m)P(s^m \mid i^0) = 0.927 \cdot 0.95 + 0.035 \cdot 0.05 = 0.8824
\]

In general, if \( A_x \) and \( B_x \) are subsets of domain \( \Omega_X \), then we cannot insert the unit operator \( I_y \) into their conditional bracket:

\[
P(A_x \mid B_x) = P_{X \mid x}(A_x \mid B_x) \neq P(A_x \mid I_y \mid B_x) \tag{2.2.31}
\]

Because observable \( Y(X) \) has no fixed values for the \( P \)-basis of domain \( \Omega_X(\Omega_Y) \), the following expressions are meaninglessness:

\[
P(A_x \mid Y \mid x), \quad A_x \subseteq \Omega_X; \quad P(B_y \mid X \mid y), \quad B_y \subseteq \Omega_Y \tag{2.2.32}
\]

On the other hand, the following cross-domain expression is well-defined:

\[
P(A_x \mid Y \mid y) = yP(A_x \mid y) = y\sum_x P(A_x \mid x)P(x \mid y) = y\sum_{x \in A_x} P(x \mid y) \tag{2.2.33}
\]

\textbf{Note -2.2.1:} As one can see from Eq. (2.2.18), (2.2.26) and (2.2.29), JPD, IPD and MPD of a multi-variable system are cross-domain brackets by nature.

\textbf{Note-2.2.2:} We will not discuss continuous variables, but related expressions can be easily obtained from those of discrete variables [1] [7]. For example, if \( X \) and \( Y \) both are
continuous *real* variables, then the orthonormality and completeness in Eq. (2.2.2) become:

\[ P(x, y \mid x', y') = \delta(x - x')\delta(y - y'), \quad \int_{\Omega} dx\, dy \mid x, y)P(x, y) = I_{x,y} \]  

(2.2.34)

And the CP normalization (2.2.19a) becomes:

\[ 1 = P(\Omega \mid x) = P(\Omega \mid I_y \mid x) = \int dy\, P(\Omega \mid y)P(y \mid x) = \int dy\, P(y \mid x) \]  

(2.2.36)

### 2.3 Three-Variable Systems and the Student D-I-G Example

Suppose we have three discrete random variables \( X, Y \) and \( Z \). Their sample spaces are \( \Omega_x, \Omega_y \) and \( \Omega_z \) respectively and we also have a joint sample space \( \Omega = \{\Omega_x, \Omega_y, \Omega_z\} \), in which we have the following \( P \)-basis for any event \( \{x, y, z\} \in \{\Omega_x, \Omega_y, \Omega_z\} \):

\[ X \mid x, y, z = x \mid x, y, z, \quad Y \mid x, y, z = y \mid x, y, z, \quad Z \mid x, y, z = z \mid x, y, z \]  

(2.3.1a)

\[ P(\Omega \mid x, y, z) = 1 \]  

(2.3.1b)

Orthonormality: \( P(x, y, z \mid x', y', z') = \delta_{xx'}\delta_{yy'}\delta_{zz'} \)  

(2.3.2)

Completeness: \( \sum_{x,y,z \in \Omega} |x, y, z)P(x, y, z \mid = I_{x,y,z} \)  

(2.3.3)

The *full joint probability distribution* (JPD) is defined by:

\[ P(x, y, z) = P(x, y, z \mid \Omega) = P(X = x, Y = y, Z = z \mid \Omega), \quad \{x, y, z\} \in \Omega \]  

(2.3.4)

It has following property:

\[ P(x, y, z) \begin{cases} \geq 0, & \text{if } \{x, y, z\} \in \Omega \\ = 0, & \text{if } \{x, y, z\} \notin \Omega \end{cases} \]  

(2.3.5)

In addition to the properties and definitions we have seen for two observables, we have CPD related to three observables. For example, here is the CPD over \( X \) and \( Y \) given \( Z \):

\[ P(x, y \mid z) \equiv P(X = x, Y = y \mid Z = z) = P_{X,Y \mid Z}(x, y \mid z) \]  

(2.3.6)

**Definition 2.3.1:** Event \( X = x \) is *conditionally independent* of event \( Y = y \) given \( Z = z \) in a distribution \( P \), denoted \( P \vdash (X \perp Y \mid z) \), if \( P(x \mid y \wedge z) \equiv P(x \mid y, z) = P(x \mid z) \) or if \( P(y \wedge z) \equiv P(y, z) = 0 \) (§2.4.2 of [3]).

(2.3.7)

**Definition 2.3.2:** Random variables \( X \) and \( Y \) are conditional independent given variable \( Z \) in a distribution \( P \), denoted \( P \vdash (X \perp Y \mid Z) \), if \( P \vdash (x \perp y \mid z) \) for \( \forall x, y, z \in \Omega \).  

(2.3.8)
Proposition 2.3.1: \( P \models (X \perp Y \mid Z) \) if and only if \( P(X, Y \mid Z) = P(X \mid Z)P(Y \mid Z) \). (2.3.9)

Proof: see §2.1.4.2 of [3].

If \( P \models (X \perp Y \mid Z) \), then using Bayes’ rule and Eq. (2.3.9), we have the following JPD:

\[
P(x, y, z) = P(x, y \mid z)P(z) = P(x \mid z)P(y \mid z)P(z)
\]

It is the product of the PDF in following graphical representation (Fig.2.3.1):

\[
\text{Fig. 2.3.1: } P \models (X \perp Y \mid Z)
\]

Fig 2.3.2 shows another conditional independence of a 3-variable system.

\[
\text{Fig. 2.3.2: } Y \text{ depends on } X \text{ and } Z \text{ depends on } Y
\]

From Fig. 2.3.2 and the chain rule Eq. (1.3.4) we have following two JPD respectively:

\[
P(x, y, z) = P(x)P(y \mid x)P(z \mid y)
\]

(2.3.11)

\[
P(x, y, z) = P(x)P(y \mid x)P(z \mid x, y)
\]

(2.3.12)

Comparing Eq. (2.3.11) and (2.3.12), we get:

\[
P(z \mid x, y) = P(z \mid y) \Rightarrow P \models (Z \perp X \mid Y)
\]

(2.3.13)

Now let us discuss the \textbf{D-I-G example of the student Bayesian network} [3], which has three variables: \textit{D} (the difficulty of a course), \textit{I} (the intelligence of a student) and \textit{G} (the student’s grade of the course), as show in Fig. 2.3.3:
The primary MPD of $P(I)$ and $P(D)$ as well as the CPD of $P(F \mid I, D)$ are given in Fig. 3.1.2 (on page 17).

From Fig. 2.3.3 and the chain rule Eq. (1.3.4) we have following two JPD respectively:

$$P(d, i, g) = P(d)P(i)P(g \mid d, i)$$  \hspace{1cm} (2.3.14)
$$P(d, i, g) = P(d)P(i \mid d)P(g \mid d, i)$$  \hspace{1cm} (2.3.15)

Comparing Eq. (2.3.14) and (2.3.15), we find:

$$P(i \mid d) = P(i) \Rightarrow P(\mid (D \perp I) \text{ or } P(d, i) = P(d)P(i)$$  \hspace{1cm} (2.3.16)

From Eq. (2.3.14), we can calculate miscellaneous intermediate JPD and MPD, like:

$$P(d, g) = \sum_i P(d, i, g) = \sum_i P(d)P(i)P(g \mid d, i)$$  \hspace{1cm} (2.3.17)
$$P(d) = \sum_g P(d, g) = \sum_{i,g} P(d)P(i)P(g \mid d, i) = \sum_i P(d)P(i) = P(d)$$  \hspace{1cm} (2.3.18)

Here we have used the CP normalization similar to Eq. (2.2.19):

$$1 = P(\Omega \mid d, i) = \sum_g P(\Omega \mid I_G \mid d, i) = \sum_g P(g \mid d, i)$$  \hspace{1cm} (2.3.19)

Note that we can derive $P(d, g)$ by using the unit operator $I_i$ in Eq. (2.2.26):

$$P(d, g) = P(d, g \mid \Omega) = \sum_i P(d, g \mid i)P(i \mid \Omega) = \sum_i \frac{P(d, i, g)}{P(i)}P(i) = \sum_i P(d, i, g)$$  \hspace{1cm} (2.3.20)

Similarly, we can derive $P(i, d)$ by using unit operator $I_D$:

$$P(i, g) = P(i, g \mid \Omega) = P(i, g \mid I_D \mid \Omega) = \sum_d P(i, g \mid d)P(d \mid \Omega)$$
$$= \sum_g \frac{P(d, i, g)}{P(d)}P(d) = \sum_d P(d, i, g)$$  \hspace{1cm} (2.3.21)
The MPD of $G$ can be derived from the definition and the chain rule:

$$P(g) = \sum_{i,d} P(g, i, d) = \sum_{i,d} P(g \mid i, d) P(i) P(d)$$  \hspace{1cm} (2.3.22)

Alternatively, we can also derive it by inserting unit operator $I_{i,d}$:

$$P(g) = P(g \mid I_{i,d} \mid \Omega) = \sum_{i} P(g \mid i, d) P(i, d \mid \Omega) = \sum_{i} P(g \mid i, d) P(i) P(d)$$  \hspace{1cm} (2.3.23)

Now we are able to derive various CPD from Eq. (2.3.17-23), like:

$$P(i \mid g) = \frac{P(i, g)}{P(g)} = \frac{1}{P(g)} \sum_{d} P(i, d, g) = \frac{1}{P(g)} \sum_{d} P(g \mid i, d) P(i) P(d)$$  \hspace{1cm} (2.3.24)

Note: although $D$ and $I$ are independent if $G$ is not given, they are correlated if $G$ is given (see Fig. 3.3.4). The $D-I-G$ BN in Fig. 2.3.3 is said to have a **V-shape structure** [5].

### 2.4 Properties of Conditional Independence

The 5 most important properties are listed here, as in Lecture 2 of Ref. [6].

1. **Symmetry**
   $$X \perp Y \mid Z \Rightarrow Y \perp X \mid Z$$  \hspace{1cm} (2.4.1)

2. **Decomposition**
   $$X \perp \{Y, W\} \mid Z \Rightarrow X \perp Y \mid Z$$  \hspace{1cm} (2.4.2)

3. **Contraction**
   $$(X \perp Y \mid Z) \land (X \perp W \mid Y, Z) \Rightarrow X \perp \{Y, W\} \mid Z$$  \hspace{1cm} (2.4.3)

4. **Weak union**
   $$X \perp \{Y, W\} \mid Z \Rightarrow X \perp Y \mid Z, W$$  \hspace{1cm} (2.4.4)

5. **Intersection**
   $$(X \perp Y \mid Z, W) \land (X \perp W \mid Y, Z) \Rightarrow X \perp \{Y, W\} \mid Z$$  \hspace{1cm} (2.4.5)

Holds only if distribution is positive, i.e., $P > 0$

The proof of (2.4.2) can be found in Ref [3]. Here we give a proof of Weak union:

Suppose $X \perp \{Y, W\} \mid Z$ is true, then: $P(X, Y, W \mid Z) = P(X \mid Z) P(Y, W \mid Z)$

$$\therefore P(X, W \mid Z) = \sum_{y} P(X, y, W \mid Z) = \sum_{y} P(X \mid Z) P(y, W \mid Z) = P(X \mid Z) P(W \mid Z)$$  \hspace{1cm} (2.4.7)
And: \( P(X | W, Z) = \frac{P(X, Z, W)}{P(W, Z)} = \frac{P(X, W | Z)P(Z)}{P(W | Z)} = \frac{P(X, W | Z)}{P(W | Z)} = P(X | Z) \) \hspace{1cm} (2.4.8)

Therefore: \( P(X, Y | W, Z) = \frac{P(X, Y, W, Z)}{P(W, Z)} = \frac{P(X, Y, W | Z)P(Z)}{P(W, Z)} = \frac{P(X | Z)P(Y, W, Z)P(Z)}{P(Z)P(W, Z)} = P(X | Z)P(Y | W, Z) \)

\( = \frac{P(X | W, Z)P(Y | W, Z)}{P(W, Z)} \) \hspace{1cm} (2.4.9)

We actually have shown: \( X \perp \{Y, W\} | Z \Rightarrow (X \perp W | Z) \land (X \perp Y | Z, W) \) \hspace{1cm} (2.4.10)

### 2.5 Definitions and Rules on Bayesian Networks (Informal)

Here we introduce some informal graphic definitions and rules for BN. Please see Ref [3] [4] [5] for more rigorous details.

The following informal definitions and rules are quoted from Lecture 2 of Ref. [5].

**Bayesian network (informal)**

- Directed acyclic graph (DAG) \( G \)
- Nodes represent random variables
- Edges represent direct influences between random variables
- Local probability models (conditional parameterization)
- Conditional probability distributions (CPDs)

**Bayesian network structure**

- Directed acyclic graph (DAG) \( G \)
- Nodes \( X_1, \ldots, X_n \) represent random variables
- \( G \) encodes the following set of independence assumptions (called local independencies)
  - \( X_i \) is independent of its non-descendants given its parents
  - Formally: \( (X_i \perp NonDesc(X_i) | P_{s}(X_i)) \)
  - Denoted by \( I_s(G) \)

### 3. The Student Bayesian Network and Reasoning Examples

Now we are ready to study the famous Student Bayesian Network [3-5]. In §3.1, we will show its GPM, local dependences, the probability distribution function (PDF) for each
node and some reasoning results. Next, in §3.2, we will do our homework to derive these reasoning results, and then, in §3.3, we will show related screenshots by using Elvira.

3.1 Local Independences of Student BN and Some Reasoning Results

The Student BN [3-5] is graphically shown in Fig. 3.1.1 below.

Here we have 5 random variables with their sample spaces \( \{L, D, I, G, S\} \) or \( \Omega_D, \Omega_I, \Omega_G, \Omega_S \) and \( \Omega_L \) as shown in Fig. 3.1.1. Applying the rules in §2.5 we have the following local independences [5] for SBN:

\[
I_L(G_{\text{student}}):
\begin{align*}
D \perp I & \quad \text{(3.1.1a)} \\
D \perp S & \quad \text{(3.1.1b)} \\
G \perp S | D, I & \quad \text{(3.1.1c)} \\
L \perp I, D, S | G & \quad \text{(3.1.1d)} \\
S \perp D, G, L | I & \quad \text{(3.1.1e)}
\end{align*}
\]

![Fig. 3.1.1: The Student Bayesian Network Example (see Ref [5])](image)

To be consistent with [3], we list the values and associated PDF of the 5 nodes of SBN in Fig. 3.1.1 as follows:

Intelligence: \( \text{Val}(I) = \{i^0, i^1\} = \{\text{low, high}\} = \Omega_I \) \( (3.1.2a) \)
Difficulty: \( \text{Val}(D) = \{d^0, d^1\} = \{\text{easy, hard}\} = \Omega_D \)  
(3.1.2b)

Grade: \( \text{Val}(G) = \{g^1, g^2, g^3\} = \{A, B, C\} = \Omega_G \)  
(3.1.2c)

SAT: \( \text{Val}(S) = \{s^0, s^1\} = \{\text{low, high}\} = \Omega_S \)  
(3.1.2d)

Letter: \( \text{Val}(L) = \{l^0, l^1\} = \{\text{weak, strong}\} = \Omega_L \)  
(3.1.2e)

The GPM in Fig 3.1.1 tell us about the PDF associated with each node:

\{D, I, G, S, L\} \Rightarrow \{P(d), P(i), P(g \mid d, i), P(s \mid i), P(l \mid g)\} \tag{3.1.3}

It also tells us the expression of the full joint probability distribution:

\[ P(d, i, g, s, l) = P(d)P(i)P(g \mid d, i)P(s \mid i)P(l \mid g) \] \tag{3.1.4}

On the other hand, using the chain rule Eq. (1.3.2), we have

\[ P(d, i, g, s, l) = P(d)P(i \mid d)P(g \mid d, i)P(s \mid d, i, g)P(l \mid d, i, g, s) \] \tag{3.1.5}

We can use relations in Eq. (3.1.1) to derive Eq. (3.1.5) from Eq. (3.1.4). But, as one can see, it is much easier to get Eq. (3.15 by using PGM for BN.

The numerical values of the PDF tables for the 5 nodes in Eq. (3.1.3) are given in Fig.3.1.2 below (from Ref [4], corrected according to [3]).

---

**Fig. 3.1.2: The PDF for Student BN ([3], [4])**
To simplify our statements, we assume the student has name George and the course name is Econ101 [3]. Then, from student BN, we can make **direct predictions** like:

1. If George has grade \( g^1 = A \), then the probability for him to get a strong letter from his professor is \( P(l^1 \mid g^1) = 0.9 \), independent of the values of \( D, I \) or \( S \).
2. If George has high intelligence \( (i^1) \) and the course is hard \( (d^1) \), then the probability for him to get grade \( g^1 = A \) is \( P(g^1 \mid i^1, d^1) = 0.5 \), independent of \( S \).
3. If George has low intelligence \( (i^0) \), then the probability for him to get high SAT \( (s^1) \) is \( P(s^1 \mid i^0) = 0.05 \), independent of \( D, G \) or \( L \).

Another important aspect of BN is its power of **reasoning** (top-down or bottom-up, prediction, explaining-away or inference [3]). Here are some reasoning examples of SBN, many of which are given in Ref. [3] without derivation. In next section, we will treat them as **our homework**, to derive them based on the PDF tables given in Fig. 3.1.2.

**HW-1:** Assuming we know nothing about Econ101 and George, \( P(I) \) in Fig. 3.1.2 tells us that, George has 30% chance to have high intelligence, or \( P(i^1) = 0.3 \). But, if we know his grade of Econ101 is \( g^3 = C \), then his chance of high intelligence is reduced to 7.9%: \( P(i^1 \mid g^3) = 0.079 \) (as a case of inference in bottom-up reasoning).

**HW-2:** Assuming we know nothing about Econ101 and George, one can find that he will have a chance of 36.2% to get a grade \( g^1 = A \), or \( P(g^1) = 0.362 \). But, if we know that George has low intelligence \( (i^0) \), then the chance for him to get grade \( A \) is reduced to 20%: \( P(g^1 \mid i^0) = 0.20 \) (as a case of prediction in top-down reasoning).

**HW-3:** Assuming we know nothing about Econ101 and George, one can find that he will have 50.2% chance to get a strong recommendation letter, or \( P(l^1) = 0.502 \). But, if George has low intelligence \( (i^0) \), then his chance to get a strong letter will be reduced to 38.9%, or \( P(l^1 \mid i^0) = 0.389 \). Now if we also know that the course is easy, then his chance to get a strong letter will be increased to 51.3%: \( P(l^1 \mid i^0, d^0) = 0.513 \) (as a case of explaining-away in top-down reasoning).

### 3.2 Homework Exercises: Deriving the Reasoning Results for Student BN

To do our home work, we need to use miscellaneous marginal and conditional PDF, defined in or to be derived from the tables in Fig. 3.1.2.

**3.2.1: The Marginal PDF of \( G \) (Grade) or \( P(g) = P(g \mid \Omega) \):**
Because we know $P(g \mid i, d)$ and $P(i, d \mid \Omega) = P(i, d) = P(i)P(d)$, it is naturally to insert unit operator $I_{i,d}$, as we did in Eq. (2.3.25):

$$P(g) = P(g \mid \Omega) = P(g \mid I_{i,d} \mid \Omega) = \sum_{i,d} P(g \mid i, d)P(i, d)$$

$$= \sum_{i,d} P(g \mid i, d)P(i)P(d)$$

(3.2.1)

This is an alternative way to derive Eq. (3.2.1), which usually is obtained from the joint PDF and chain rule:

$$P(g) = \sum_{i,d} P(g, i, d) = \sum_{i,d} P(g \mid i, d)P(i, d)$$

It is now straightforward to evaluate $P(g)$:

$$P(g^\mu) = \sum_{r,\rho} P(g^\mu \mid i^r, d^\rho)P(i^r)P(d^\rho)$$

$$P(g^\mu) = P(g^\mu \mid i^0, d^0)P(i^0)P(d^0) + P(g^\mu \mid i^0, d^1)P(i^0)P(d^1)$$

$$+ P(g^\mu \mid i^1, d^0)P(i^1)P(d^0) + P(g^\mu \mid i^1, d^1)P(i^1)P(d^1)$$

(3.2.2)

Using tables in Fig. 3.1.2 and Eq. (3.2.2), we have:

- $P(g^1) = 0.3620$ (as in HW-2)  
  \hspace{1cm} (3.2.3a)
- $P(g^2) = 0.2884$  
  \hspace{1cm} (3.2.3b)
- $P(g^3) = 0.3496$  
  \hspace{1cm} (3.2.3c)

One can check that the marginal normalization is valid:

$$\sum_{\mu} P(g^\mu) = 1.000$$

(3.2.4)

**3.2.2: The Marginal PDF of $L$ (Letter) or $P(l) = P(l \mid \Omega)$:** Because we already know $P(l \mid g)$ and $P(g \mid \Omega) = P(g)$, it is naturally to insert unit operator $I_g$:

$$P(l) = P(l \mid \Omega) = P(l \mid I_g \mid \Omega) = \sum_g P(l \mid g)P(g \mid \Omega) = \sum_g P(l \mid g)P(g)$$

(3.2.5)

Equivalently, it can be derived by using Bayes’ chain rule as:

$$P(l) = \sum_g P(l, g) = \sum_g P(l \mid g)P(g)$$

(3.2.6)

Using tables in Fig. 3.1.2 and Eq. (3.2.6), it is straightforward to calculate:
\( P(l^i) \approx 0.502 \) (as in HW-3) \hspace{1cm} (3.2.7)

**3.2.3: The Conditional PDF of \( I \) (Intelligence) given \( G \) or \( P(i \mid g) \):**

The expression has already derived in Eq. (2.3.24):

\[
P(i \mid g) = \frac{P(i, g)}{P(g)} = \frac{1}{P(g)} \sum_d P(g \mid i, d)P(i)P(d)
\]  \hspace{1cm} (3.2.8)

Using tables in Fig. 3.1.2 and Eq. (3.2.3), it is easy to obtain:

\( P(i^1 \mid g^3) \approx 0.079 \) (as in HW-1) \hspace{1cm} (3.2.9)

**3.2.4: The Conditional PDF of \( L \) (Letter) given \( I \) and \( D \) or \( P(l \mid i, d) \):**

\[
P(l \mid i, d) = P(l \mid I_G, i, d) = \sum_g P(l \mid g)P(g \mid i, d)
\]  \hspace{1cm} (3.2.10)

Using tables in Fig. 3.1.2, one gets:

\( P(l^1 \mid i^0, d^0) \approx 0.513 \) (as in HW-3) \hspace{1cm} (3.2.11)

**3.2.5: The Conditional PDF of \( G \) (Grade) given \( I \) or \( P(g \mid i) \):**

\[
P(g \mid i) = \frac{P(g, i)}{P(i)} = \sum_d \frac{P(g, i, d)}{P(i)} = \sum_d \frac{P(g \mid i, d)P(i, d)}{P(i)} = \sum_d P(g \mid i, d)P(d)
\]  \hspace{1cm} (3.2.12)

Assuming that George has low intelligence and using tables in Fig. 3.1.2, we have:

\( P(g^1 \mid i^0) = 0.20 \) (as in HW-2) \hspace{1cm} (3.2.13a)
\( P(g^2 \mid i^0) = 0.34 \) \hspace{1cm} (3.2.13b)
\( P(g^3 \mid i^0) = 0.46 \) \hspace{1cm} (3.2.13c)

Again, the CP normalization is satisfied:

\( \sum_g P(g \mid i^0) = 1.00 \) \hspace{1cm} (3.2.14)

**3.2.6: The Conditional PDF of \( L \) (Letter) given \( I \) or \( P(l \mid i) \):**
\[ P(I \mid i) = P(I \mid I_g \mid i) = \sum_g P(I \mid g)P(g \mid i) \quad (3.2.15) \]

Using table \( P(I \mid g) \) in Fig 3.1.2 and Eq. (3.2.13), we obtain:

\[ P(I^l \mid i^0) = \sum_g P(I \mid g)P(g \mid i^0) \approx 0.389 \quad \text{(as in HW-3)} \quad (3.2.16) \]

Thus we have completed our homework exercises for student BN, as mentioned in §3.1.

As we have shown, many probabilistic expressions can be obtained by simply inserting appropriate unit operators into appropriate (cross-domain) probability brackets. Therefore, \textit{PBN} seems to be able to provide us with a very useful alternative in manipulating various probability distributions for multivariable systems.

### 3.3 Graphic presentation of Student Bayesian Network by Using Elvira

There are various software packages for building and testing BN [8]. Here we give four screenshots of student BN by using \textit{Elvira} [9], which is a Java application with a very nice graphic user interface. We find it is a great tool to be used to learn BN and PGM.

![Fig. 3.3.1: Student BN (in Elvira Edit mode showing link inferences)](image-url)
Fig. 3.3.1 (above): Edit mode. The data of PD tables (under tab Relation in the “Edit node properties…” dialog) for each node are from Fig. 3.1.2, with a order from low (weak, easy) to higher (strong, hard) as shown in Fig. 3.3.2 The graph also shows the link inferences: red link means positive (e.g., higher intelligence causes higher grade and SAT), while blue link means negative (e.g., the harder the course, the lower the grade).

![Fig. 3.3.1 (above): Edit mode.](image)

**Fig. 3.3.2**: Student BN (in Elvira Inference mode with prior probabilities)

**Fig. 3.3.2 (above): Inferences with prior probabilities** (without any additional information). We see that George will have 50.2% chance to get a strong recommendation letter as in Eq. (3.2.7), and his PD of grade is as shown in Eq. (3.2.3).

**Fig. 3.3.3 (below): Inferences, assuming George has low intelligence.** We see that he will have 38.9% chance to get a strong recommendation letter, as in Eq. (3.2.16), and his PD of grade is changed to Eq. (3.2.3).

**Fig. 3.3.4 (below): Inferences, assuming George has grade = C.** We see that his chance to have high intelligence reduced to 0.079, as in Eq. (3.2.9).

Thus we have confirmed our results for student BN by using Elvira. The student BN is saved as an Elvira-formatted file [10] and can be reused by clicking “Open Network…” under File menu in Elvira. If user is interested, Elvira can be easily installed, and there are many interesting (simple and complex) samples already in Elvira-format [9].
Fig. 3.3.2: Student BN (in Elvira Inference mode with Intelligence=low)

Fig. 3.3.4: Student BN (in Elvira Inference mode with Grade=C)
Summary

We applied PBN to study discrete multivariable systems in static Bayesian Networks (BN). We found that, by inserting appropriate unit operators into appropriate (cross-domain) probability brackets, PBN gave us a consistent and convenient shortcut to manipulate miscellaneous relations between joint, intermediate, marginal or conditional probability distributions (JPD, IPD, MPD or CPD). We demonstrated it further by doing our homework related to the local independences and reasoning power of the famous student BN example. In last section, we used Elvira, a Java application, to graphically display and verify our inferences with the student BN example.

The goal of this article is not to formally introduce static BN and probabilistic graphic models (PGM). Readers may exploit the excellent textbook [3] and many good lecture notes online [4-6] for that purpose. In this article, we just want to show that, PBN, as an alternative tool, can also be applied to study dependent variables in static BN.

References

[1]. Wang X, Probability Bracket Notation, Probability Vectors, Markov Chains and Stochastic processes, see http://arxiv.org/abs/cs.IR/0702021
[2]. Wikipedia, Joint probability distribution and Bayesian network, see http://en.wikipedia.org/wiki/Joint_probability_distribution and http://en.wikipedia.org/wiki/Bayesian_network
[3]. Koller D. and Friedman N, Probabilistic Graphic Models. MIT Press, 2009.
[4]. Sontag D, Probabilistic Graphic Models, Lecture 1. See: http://cs.nyu.edu/~dsontag/courses/pgm12/slides/lecture1.pdf
[5]. Lee S, Statistical Methods in Computer Science, Lectures of CSE 515, see http://www.cs.washington.edu/education/courses/cse515/11sp/lectures.html
[6]. Krause A, Probabilistic Graphic Models, Lectures of CS 155, see http://courses.cms.caltech.edu/cs155/
[7]. Wang X, Probability Bracket Notation: the Unified Expressions of Conditional Expectation and Conditional Probability in Quantum Modeling, see: http://arxiv.org/abs/0911.1462
[8]. Murphy K, see: http://www.cs.ubc.ca/~murphyk/Software/bnsoft.html
[9]. Elvira System Home Page, see: http://leo.ugr.es/elvira/elvira.html and http://www.ia.uned.es/~elvira/index-en.html
[10]. Wang X, StudentBayesNet.elv, please view the page source of: http://www.shermanlab.com/science/CS/AI/StudentBayesNet.elv