Lecture 8 - Reasoning under Uncertainty (Part I)

Jesse Hoey
School of Computer Science
University of Waterloo

June 2, 2022

Readings: Poole & Mackworth (2nd ed.) Chapt. 8 up to 8.4
Why is uncertainty important?

- Agents (and humans) don’t know \textit{everything},
- but need to \textit{make decisions} anyways!
- Decisions are made in the \textit{absence of information},
- or in the presence of \textit{noisy} information (sensor readings)

The best an agent can do:
- \textit{know} how uncertain it is, and act accordingly
Probability: Frequentist vs. Bayesian

**Frequentist view:**
probability of heads = # of heads / # of flips
probability of heads *this time* = probability of heads (history)
Uncertainty is **ontological**: pertaining to the world

**Bayesian view:**
probability of heads *this time* = agent’s **belief** about flip
belief of agent A: based on **previous experience** of agent A
Uncertainty is **epistemological**: pertaining to knowledge
Bayesian probability
all else being equal (Prior)
before any flips
Bayesian probability
all else being equal (Prior)
after 2 flips **heads, heads** (Posterior)
Bayesian probability
all else being equal (Prior)
after 2 flips tails,tails (Posterior)
DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)

This neutrino detector measures whether the sun has gone nova.

Then, it rolls two dice. If they both come up six, it lies to us. Otherwise, it tells the truth.

Let's try.
Detector! Has the sun gone nova?

Roll
Yes.

Frequentist Statistician:
The probability of this result happening by chance is \( \frac{1}{36} = 0.027 \).
Since \( p < 0.05 \), I conclude that the sun has exploded.

Bayesian Statistician:
Bet you $50 it hasn't.
Probability: Bayesian

Should you **wear your seatbelt**? estimate \( P(\text{injury}) \) given you do/don’t wear it
Probability: Bayesian

Should you **wear your seatbelt**?

estimate \( P(\text{injury}) \) given you do/don’t wear it

**Frequentist:**

<table>
<thead>
<tr>
<th>test</th>
<th>day</th>
<th>result</th>
<th>( P(\text{fatality}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Monday</td>
<td><img src="#" alt="Car" /></td>
<td>0.0</td>
</tr>
<tr>
<td>2</td>
<td>Tuesday</td>
<td><img src="#" alt="Car" /></td>
<td>0.0</td>
</tr>
<tr>
<td>3</td>
<td>Tuesday</td>
<td><img src="#" alt="Car, Truck" /></td>
<td>0.33333</td>
</tr>
<tr>
<td>4</td>
<td>Thursday</td>
<td><img src="#" alt="Car" /></td>
<td>0.25</td>
</tr>
<tr>
<td>5</td>
<td>Friday</td>
<td><img src="#" alt="Car" /></td>
<td>0.2</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>N</td>
<td><strong>Number of injuries / N</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Should you wear your seatbelt?

estimate $P(\text{injury})$ given you do/don’t wear it

Bayesian:

UK government: "Seatbelts save 2200 lives/year"

car-accidents.com: "in 63% of fatalities, no seatbelts were worn"
if $X$ is a random variable (feature, attribute),
it can take on values $x$, where $x \in \text{Domain}(X)$ (or $\text{Dom}(X)$)
Assume $x$ is discrete

$P(x)$ is the probability that $X = x$

**joint probability** $P(x, y)$ is the
probability that $X = x$ and $Y = y$ at the same time

Joint probability distribution:

```
   1  2  3  4  5  6  7  8  9
10  ...  ...
   0.1
   0.1  0.7  0.05
   0.01
   0.01  0.02
   0.01
```

Where is the robot?
features: $X,Y$
Axioms of Probability

Axioms are things we have to assume about probability:

- \( P(X) \geq 0 \)
- \( \sum_x P(X = x) = 1.0 \)
- \( P(a \lor b) = P(a) + P(b) \) if \( a \) and \( b \) are contradictory - can’t both be true at the same time e.g. \( P(\text{win} \lor \text{lose}) = P(\text{win}) + P(\text{lose}) = 1.0 \)

Some notes:

- probability between 0-1 is purely convention
- \( P(a) = 0 \) means you think \( a \) is definitely false
- \( P(a) = 1 \) means you think \( a \) is definitely true
- \( 0 < P(a) < 1 \) means you have belief about the truth of \( a \). It does not mean that \( a \) is true to some degree, just that you are ignorant of its truth value.
- Probability = measure of ignorance
Independence

- describe a system with \( n \) features: \( 2^n - 1 \) probabilities
- Use **independence** to reduce number of probabilities
- e.g. radially symmetric dartboard, \( P(\text{hit a sector}) \)
- \( P(\text{sector}) = P(r, \theta) \) where \( r = 1, \ldots, 4 \) and \( \theta = 1, \ldots, 8 \).
- 32 sectors in total - need to give 31 numbers
Independence

- describe a system with $n$ features: $2^n - 1$ probabilities
- Use independence to reduce number of probabilities
- e.g. radially symmetric dartboard, $P$(hit a sector)
- assume radial independence: $P(r, \theta) = P(r)P(\theta)$
- only need $7 + 3 = 10$ numbers
Independence

- describe a system with $n$ features: $2^n - 1$ probabilities
- Use **independence** to reduce number of probabilities
- e.g. radially symmetric dartboard, $P($hit a sector$)$
- assume **radial independence** : $P(r, \theta) = P(r)P(\theta)$
- only need $7+3=10$ numbers
if $X$ and $Y$ are random variables, then

$$P(x|y)$$ is the probability that $X = x$ given that $Y = y$.

e.g. $P(flies|is\_bird)$ is different than $P(flies)$
$P(flies|is\_a\_penguin, is\_bird)$ is different again

**incorporate independence:**

$P(flies|is\_bird, has\_feathers) = P(flies|is\_bird)$

**Product rule (Chain rule):**

$P(flies, is\_bird) = P(flies|is\_bird)P(is\_bird)$

$P(flies, is\_bird) = P(is\_bird|flies)P(flies)$

leads to: Bayes’ rule

$$P(is\_bird|flies) = \frac{P(flies|is\_bird)P(is\_bird)}{P(flies)}$$
Sum Rule

We know (an Axiom):

$$\sum_x P(X = x) = 1.0$$ and therefore that $$\sum_x P(X = x \mid Y) = 1.0$$

This means that (Sum Rule)

$$\sum_x P(X = x, Y) = P(Y)$$

proof:

$$\sum_x P(X = x, Y) = \sum_x P(X = x \mid Y)P(Y)$$

$$= P(Y) \sum_x P(X = x \mid Y)$$

$$= P(Y)$$

We call $P(Y)$ the marginal distribution over $Y$
Conditional Probability

- **X** and **Y** are **independent** iff

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

so learning **Y** doesn’t influence beliefs about **X**

- **X** and **Y** are **conditionally independent** given **Z** iff

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

so learning **Y** **doesn’t influence beliefs** about **X** if you already know **Z**...does **not** mean **X** and **Y** are independent
Expected Values

**expected value** of a function on $X$, $V(X)$:

$$
\mathbb{E}(V) = \sum_{x \in \text{Dom}(X)} P(x) V(x)
$$

where $P(x)$ is the probability that $X = x$.

This is useful in **decision making**, where $V(X)$ is the *utility* of situation $X$.

**Bayesian decision making** is then

$$
\mathbb{E}(V(\text{decision})) = \sum_{\text{outcome}} P(\text{outcome} | \text{decision}) V(\text{outcome})
$$
complete independence reduces both **representation** and **inference** from $O(2^n)$ to $O(n)$

Unfortunately, complete mutual independence is **rare**

Fortunately, most domains do exhibit a fair amount of **conditional independence**

**Bayesian Networks** or **Belief Networks** (BNs) encode this information
Belief Networks

Bayesian network or belief network
- Directed Acyclic graph
- Encodes independencies in a graphical format
- Edges give \( P(X_i|\text{parents}(X_i)) \)

Cancer diagnosis example:
- Two tests A and B
- Test A is quick and cheap, but imprecise
- Test A results are read directly
- Test B uses a machine that sometimes malfunctions, but is more precise
- Test B results are not read directly, a Report is written (by a human who makes mistakes)
- the Report is entered into a database (by another human who makes mistakes)
Directed links in Bayes’ net \( \approx \text{causal} \)

However, \textbf{not always} the case: chocolate \( \rightarrow \) Nobel or Nobel \( \rightarrow \) chocolate?

In a Bayes net, it doesn’t matter!

But, some structures will be \textbf{easier to specify}.

In this example, it’s probably chocolate \( \leftarrow \) “Switzerland – ness” \( \rightarrow \) Nobel.
If Jesse’s alarm doesn’t go off (A), Jesse probably won’t get coffee (C); if Jesse doesn’t get coffee, he’s likely grumpy (G). If he is grumpy then it’s possible that the lecture won’t go smoothly L. If the lecture does not go smoothly then the students will likely be sad S.

A=Jesse’s alarm doesn’t go off
C=Jesse doesn’t get coffee
G=Jesse is grumpy
L=lecture doesn’t go smoothly
S=students are sad

all variables binary (true/false)
Conditional Independence

If you learned any of $A$, $C$, $G$, or $L$, would your assessment of $P(S)$ change?

- If any of these are seen to be true, you would increase $P(s)$ and decrease $P(\bar{s})$.
- So $S$ is not independent of $A$, $C$, $G$, $L$.

If you knew the value of $L$, would learning the value of $A$, $C$, or $G$ influence $P(S)$?

- Influence that these factors have on $S$ is mediated by their influence on $L$.
- Students aren’t sad because Jesse was grumpy, they are sad because of the lecture.
- Therefore, $S$ is conditionally independent of $A$, $C$, and $G$ (given $L$)
We say: \( S \) is independent of \( A, C, \) and \( G, \) given \( L \) (this is conditional independence).

Similarly, we can say:

\begin{itemize}
  \item \( S \) is independent of \( A \) and \( C, \) given \( G \)
  \item \( G \) is independent of \( A, \) given \( C \)
  \item ... 
\end{itemize}

This means that:

\begin{itemize}
  \item \( P(S|L, G, C, A) = P(S|L) \)
  \item \( P(L|G, C, A) = P(L|G) \)
  \item \( P(G|C, A) = P(G|C) \)
  \item \( P(C|A) \) and \( P(A) \) don’t “simplify”
\end{itemize}
Conditional Independence

Chain rule (product rule):

\[ P(S, L, G, C, A) = P(S|L, G, C, A)P(L|G, C, A)P(G|C, A)P(C|A)P(A) \]

Independence:

\[ P(S, L, G, C, A) = P(S|L)P(L|G)P(G|C)P(C|A)P(A) \]

So we can specify the full joint probability using the five local conditional probabilities:

\[ P(S|L), P(L|G), P(G|C), P(C|A), P(A) \]
A **Bayesian Network** (Belief Network, Probabilistic Network) or BN over variables \( \{X_1, X_2, \ldots, X_N\} \) consists of:

- a **DAG** whose nodes are the variables
- a set of **Conditional Probability tables** (CPTs) giving \( P(X_i | Parents(X_i)) \) for each \( X_i \)

**Example probability tables** for the Coffee Bayes Net:

\[
\begin{array}{c|cc}
A & P(A = \text{true}) = 0.3
\end{array}
\]

\[
\begin{array}{c|cc}
P(C = \text{true} | A) & T & 0.8 \\
P(C = \text{true} | A) & F & 0.15 \\
\end{array}
\]

\[
\begin{array}{c|cc}
P(G = \text{true} | C) & T & 1.0 \\
P(G = \text{true} | C) & F & 0.2 \\
\end{array}
\]

\[
\begin{array}{c|cc}
P(L = \text{true} | G) & T & 0.7 \\
P(L = \text{true} | G) & F & 0.2 \\
\end{array}
\]

\[
\begin{array}{c|cc}
P(S = \text{true} | L) & T & 0.9 \\
P(S = \text{true} | L) & F & 0.3 \\
\end{array}
\]
Another example quantification

Cancer diagnosis:

**Malfunction**

- $P(M = \text{true}) = 0.08$

**Test B**

- $P(C = \text{true}) = 0.32$

**Test A**

- $P(A = \text{true}|C) = \begin{array}{c|c}
C & P(A = \text{true}|C) \\
\hline
\text{t} & 0.80 \\
\text{f} & 0.15 \\
\end{array}$

**Report**

- $P(R = \text{true}|B) = \begin{array}{c|c}
B & P(R = \text{true}|B) \\
\hline
\text{t} & 0.98 \\
\text{f} & 0.01 \\
\end{array}$

**Database**

- $P(D = \text{true}|R) = \begin{array}{c|c}
R & P(D = \text{true}|R) \\
\hline
\text{t} & 0.96 \\
\text{f} & 0.001 \\
\end{array}$
The structure of the BN means that:

every $X_i$ is \textbf{conditionally independent} of all its \textbf{nondescendants} given its parents:

$$P(X_i|S, Parents(X_i)) = P(X_i|Parents(X_i))$$

for any subset $S \subseteq \text{NonDescendants}(X_i)$

The BN defines a \textbf{factorization} of the \textbf{joint probability} distribution. The joint distribution is formed by multiplying the conditional probability tables together.
Constructing belief networks

To represent a domain in a belief network, you need to consider:

- What are the **relevant variables**?
  - What will you observe? - this is the **evidence**
  - What would you like to find out? - this is the **query**
  - What other features make the model simpler? - these are the other variables

- What **values** should these variables take?

- What is the **relationship** between them? This should be expressed in terms of **local influence**.

- How does the value of each variable **depend on its parents**? This is expressed in terms of the **conditional probabilities**.
Bayesian Networks - Independence assumptions

- **Test B** depends on **Cancer** and **Malfunction**
- **Test A** depends only on **Cancer**
- **Report** depends only on **Test B**
- **Database** depends only on **Report**

What are the independencies?
Three Basic Bayesian Networks

Malfunction → Test B → Report → Database

Cancer → Test A → Report → Database
Three Basic Bayesian Networks

Test B

Report

Database

Database and Test B independent if Report is observed
Test B and Test A are independent if Cancer is observed.
Malfunction and Cancer are independent if Test B is not observed.
Three Basic Bayesian Networks

http://imgs.xkcd.com/comics/bridge.png
Three Basic Bayesian Networks...Recap

- Malfunction
- Cancer
- Test B
- Test A
- Report
- Database
Testing Independence

Given a BN, how do we determine if two variables $X, Y$ are independent (given evidence $E$)?

- **D-separation**: A set of variables $E$ d-separates $X$ and $Y$ if it blocks every undirected path in the BN between $X$ and $Y$.

- But what does block mean?
Blocked Paths

(1) \[ \text{If } Z \text{ in evidence, the path between } X \text{ and } Y \text{ blocked} \]

(2) \[ \text{If } Z \text{ in evidence, the path between } X \text{ and } Y \text{ blocked} \]

(3) \[ \text{If } Z \text{ is not in evidence and no descendent of } Z \text{ is in evidence, then the path between } X \text{ and } Y \text{ is blocked} \]
The **Markov Blanket** of a node (variable) $V$ is:

- the parents, children, and the (other) parents of children
- the minimal set of nodes that d-separates $V$ from all other variables

The joint distribution over the **Markov Blanket** allows for the computation of the distribution $P(V)$. 
D-Separations: Example

- TravelSubway and Thermometer (given no evidence)?
- TravelSubway and Thermometer (given Flu or Fever)?
- TravelSubway and Malaria (given Fever)?
- TravelSubway and Exotic Trip (given Jaundice)?
- TravelSubway and Exotic Trip (given Jaundice and Thermometer)?
- TravelSubway and Exotic Trip (given Malaria and Thermometer)?
Agent has a **prior belief** in a **hypothesis**, \( h, P(h) \),

Agent observes some **evidence** \( e \) that has a **likelihood** given the hypothesis: \( P(e|h) \).

The agent’s **posterior belief** about \( h \) after observing \( e \), \( P(h|e) \), is given by **Bayes’ Rule**:

\[
P(h|e) = \frac{P(e|h)P(h)}{P(e)} = \frac{P(e|h)P(h)}{\sum_h P(e|h)P(h)}
\]
Why is Bayes’ theorem interesting?

- Often you have **causal knowledge**:
 \[
P(\text{symptom} \mid \text{disease})
\]
  \[
P(\text{light is off} \mid \text{status of switches and switch positions})
\]
  \[
P(\text{alarm} \mid \text{fire})
\]
  \[
P(\text{image looks like } \text{tree} \mid \text{a tree is in front of a car})
\]

- and want to do **evidential reasoning**:
  \[
P(\text{disease} \mid \text{symptom})
\]
  \[
P(\text{status of switches} \mid \text{light is off and switch positions})
\]
  \[
P(\text{fire} \mid \text{alarm})
\]
  \[
P(\text{a tree is in front of a car} \mid \text{image looks like } \text{tree})
\]
Before you get any information

- \( P(\text{Cancer}) = 0.32 \)
- \( P(\text{Malfunction}) = 0.08 \)
Suppose the doctor reads a positive Test B in the Database evidence gives Database=true (not directly Test B= true) we want to know \( P(\text{Cancer} = \text{true}|\text{Database} = \text{true}) \)

\[ P(\text{Cancer} = \text{true}|\text{Database} = \text{true}) = 0.80 \]

\[ P(\text{Malfunction} = \text{true}|\text{Database} = \text{true}) = 0.14 \]

(we will see how to get these numbers later)
Suppose **Test A is positive** as well we want \( P(\text{Cancer} = \text{true} | \text{Database} = \text{true} \land \text{TestA} = \text{true}) \)

- \( P(\text{Cancer} = \text{true} | \text{Database} = \text{true} \land \text{TestA} = \text{true}) = 0.95 \)
- \( P(M = \text{true} | \text{Database} = \text{true} \land \text{TestA} = \text{true}) = 0.08 \)

(we will see how to get these numbers later)
Suppose \textbf{Test A is negative}, though!

we want \( P(Cancer = \text{true} | Database = \text{true} \land TestA = \text{false}) \)

\begin{itemize}
  \item \( P(Cancer = \text{true} | Database = \text{true} \land TestA = \text{false}) = 0.48 \)
  \item \( P(M = \text{true} | Database = \text{true} \land TestA = \text{false}) = 0.27 \)
\end{itemize}

(we will see how to get these numbers later)
Simple Forward Inference (Chain)

Computing marginal requires simple forward propagation of probabilities

- \( P(J) = \sum_{M,ET} P(J, M, ET) \)  
  (marginalisation - sum rule)

- \( P(J) = \sum_{M,ET} P(J|M, ET)P(M|ET)P(ET) \)  
  (chain rule)

- \( P(J) = \sum_{M,ET} P(J|M)P(M|ET)P(ET) \)  
  (conditional indep).

- \( P(J) = \sum_M P(J|M) \sum_{ET} P(M|ET)P(ET) \)  
  (distribution of sum)

Note: all terms on the last line are CPTs in the BN
Note: only ancestors of J considered. Why?
Simple Forward Inference (Chain)

Same idea when evidence “upstream”

- \( P(J|et) = \sum_M P(J, M|et) \) (marginalisation)
- \( P(J|et) = \sum_M P(J|M, et) P(M|et) \) (chain rule)
- \( P(J|et) = \sum_M P(J|M) P(M|et) \) (conditional indep).
Simple Forward Inference

With multiple parents the evidence is “pooled”

\[ P(Fev) = \sum_{Flu, M, TS, ET} P(Fev, Flu, M, TS, ET) \]

\[ = \sum_{Flu, M} P(Fev | M, Flu) \left[ \sum_{TS} P(Flu | TS) P(TS) \right] \left[ \sum_{ET} P(M | ET) P(ET) \right] \]
Simple Forward Inference

also works with “upstream” evidence

\[
P(Fev|ts, \overline{m}) = \sum_{Flu} P(Fev, Flu|\overline{m}, ts)
\]
\[
= \sum_{Flu} P(Fev|Flu, ts, \overline{m})P(Flu|ts, \overline{m})
\]
\[
= \sum_{Flu} P(Fev|Flu, \overline{m})P(Flu|ts)
\]
Simple Backward Inference

When evidence is downstream of query, then we must reason "backwards". This requires Bayes’ rule

\[
P(ET|j) = \frac{P(j|ET)P(ET)}{P(J)} \propto P(j, ET)
\]

\[
= P(j|ET)P(ET)
\]

\[
= \sum_{M} P(j, M|ET)P(ET)
\]

\[
= \sum_{M} P(j|M, ET)P(M|ET)P(ET) \quad \text{(chain rule)}
\]

\[
= \sum_{M} P(j|M)P(M|ET)P(ET)
\]

normalising constant is \(\frac{1}{P(j)}\), but this can be computed as

\[
P(j) = \sum_{ET} P(ET, j)
\]
F: Bridge on Fire
C: All friends Crazy
J: All friends Jump
What is $P(F|J = true)$?

| F | C | $P(J = true|F, C)$ |
|---|---|----------------|
| $t$ | $t$ | 0.95 |
| $t$ | $f$ | 0.99 |
| $f$ | $t$ | 0.99 |
| $f$ | $f$ | 0.01 |

$$P(C = true) = 0.0001$$

$$P(F = true) = 0.1$$
intuitions above: **polytree** algorithm
works for simple networks without loops
more general algorithm: **Variable Elimination**
applies sum-out rule repeatedly
distributes sums
A **factor** is a representation of a function from a tuple of random variables into a number. We will write factor $f$ on variables $X_1, \ldots, X_j$ as $f(X_1, \ldots, X_j)$. We can assign some or all of the variables of a factor $\rightarrow$ (this is restricting a factor):

- $f(X_1 = v_1, X_2, \ldots, X_j)$, where $v_1 \in \text{dom}(X_1)$, is a factor on $X_2, \ldots, X_j$.
- $f(X_1 = v_1, X_2 = v_2, \ldots, X_j = v_j)$ is a number that is the value of $f$ when each $X_i$ has value $v_i$.

The former is also written as $f(X_1, X_2, \ldots, X_j)_{X_1 = v_1}$, etc.
Example factors - Restricting a factor

\[ r(X, Y, Z): \]

\[
\begin{array}{ccc|c}
X & Y & Z & \text{val} \\
\hline
\text{t} & \text{t} & \text{t} & 0.1 \\
\text{t} & \text{t} & \text{f} & 0.9 \\
\text{t} & \text{f} & \text{t} & 0.2 \\
\text{t} & \text{f} & \text{f} & 0.8 \\
\text{f} & \text{t} & \text{t} & 0.4 \\
\text{f} & \text{t} & \text{f} & 0.6 \\
\text{f} & \text{f} & \text{t} & 0.3 \\
\text{f} & \text{f} & \text{f} & 0.7 \\
\end{array}
\]

\[ r(X=t, Y, Z): \]

\[
\begin{array}{cc|c}
Y & Z & \text{val} \\
\hline
\text{t} & \text{t} & 0.1 \\
\text{t} & \text{f} & 0.9 \\
\text{f} & \text{t} & 0.2 \\
\text{f} & \text{f} & 0.8 \\
\end{array}
\]

\[ r(X=t, Y, Z=f): \]

\[
\begin{array}{c|c}
Y & \text{val} \\
\hline
\text{t} & 0.9 \\
\text{f} & 0.8 \\
\end{array}
\]

\[ r(X=t, Y=f, Z=f) = 0.8 \]
The **product** of factor $f_1(X, Y)$ and $f_2(Y, Z)$, where $Y$ are the variables in common, is the factor $(f_1 \times f_2)(X, Y, Z)$ defined by:

$$(f_1 \times f_2)(X, Y, Z) = f_1(X, Y)f_2(Y, Z).$$
Multiplying factors example

### $f_1$:

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>val</th>
</tr>
</thead>
<tbody>
<tr>
<td>t</td>
<td>t</td>
<td>0.1</td>
</tr>
<tr>
<td>t</td>
<td>f</td>
<td>0.9</td>
</tr>
<tr>
<td>f</td>
<td>t</td>
<td>0.2</td>
</tr>
<tr>
<td>f</td>
<td>f</td>
<td>0.8</td>
</tr>
</tbody>
</table>

### $f_2$:

<table>
<thead>
<tr>
<th>B</th>
<th>C</th>
<th>val</th>
</tr>
</thead>
<tbody>
<tr>
<td>t</td>
<td>t</td>
<td>0.3</td>
</tr>
<tr>
<td>t</td>
<td>f</td>
<td>0.7</td>
</tr>
<tr>
<td>f</td>
<td>t</td>
<td>0.6</td>
</tr>
<tr>
<td>f</td>
<td>f</td>
<td>0.4</td>
</tr>
</tbody>
</table>

### $f_1 \times f_2$:

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>val</th>
</tr>
</thead>
<tbody>
<tr>
<td>t</td>
<td>t</td>
<td>t</td>
<td>0.03</td>
</tr>
<tr>
<td>t</td>
<td>t</td>
<td>f</td>
<td>0.07</td>
</tr>
<tr>
<td>t</td>
<td>f</td>
<td>t</td>
<td>0.54</td>
</tr>
<tr>
<td>t</td>
<td>f</td>
<td>f</td>
<td>0.36</td>
</tr>
<tr>
<td>f</td>
<td>t</td>
<td>t</td>
<td>0.06</td>
</tr>
<tr>
<td>f</td>
<td>t</td>
<td>f</td>
<td>0.14</td>
</tr>
<tr>
<td>f</td>
<td>f</td>
<td>t</td>
<td>0.48</td>
</tr>
<tr>
<td>f</td>
<td>f</td>
<td>f</td>
<td>0.32</td>
</tr>
</tbody>
</table>
We can sum out a variable, say $X_1$ with domain $\{v_1, \ldots, v_k\}$, from factor $f(X_1, \ldots, X_j)$, resulting in a factor on $X_2, \ldots, X_j$ defined by:

$$
(\sum_{X_1} f)(X_2, \ldots, X_j) = f(X_1 = v_1, \ldots, X_j) + \cdots + f(X_1 = v_k, \ldots, X_j)
$$
Summing out a variable example

\[ f_3: \]
\[
\begin{array}{ccc|c}
A & B & C & \text{val} \\
\hline
\text{t} & \text{t} & \text{t} & 0.03 \\
\text{t} & \text{t} & \text{f} & 0.07 \\
\text{t} & \text{f} & \text{t} & 0.54 \\
\text{t} & \text{f} & \text{f} & 0.36 \\
\text{f} & \text{t} & \text{t} & 0.06 \\
\text{f} & \text{t} & \text{f} & 0.14 \\
\text{f} & \text{f} & \text{t} & 0.48 \\
\text{f} & \text{f} & \text{f} & 0.32 \\
\end{array}
\]

\[ \sum_B f_3: \]
\[
\begin{array}{cc|c}
A & C & \text{val} \\
\hline
\text{t} & \text{t} & 0.57 \\
\text{t} & \text{f} & 0.43 \\
\text{f} & \text{t} & 0.54 \\
\text{f} & \text{f} & 0.46 \\
\end{array}
\]
If we want to compute the posterior probability of $Z$ given evidence $Y_1 = v_1 \land \ldots \land Y_j = v_j$:

$$P(Z|Y_1 = v_1, \ldots, Y_j = v_j) = \frac{P(Z, Y_1 = v_1, \ldots, Y_j = v_j)}{P(Y_1 = v_1, \ldots, Y_j = v_j)} = \frac{P(Z, Y_1 = v_1, \ldots, Y_j = v_j)}{\sum_Z P(Z, Y_1 = v_1, \ldots, Y_j = v_j)}.$$

The computation reduces to the joint probability of $P(Z, Y_1 = v_1, \ldots, Y_j = v_j)$.

normalize at the end.
Probability of a conjunction

Suppose the variables of the belief network are $X_1, \ldots, X_n$. To compute $P(Z, Y_1 = v_1, \ldots, Y_j = v_j)$, we sum out the variables other than query $Z$ and evidence $Y_1, \ldots, Y_j$. Let $Z_1, \ldots, Z_k = \{X_1, \ldots, X_n\} - \{Z\} - \{Y_1, \ldots, Y_j\}$. We order the $Z_i$ into an elimination ordering $Z_1 \ldots Z_k$.

$$P(Z, Y_1 = v_1, \ldots, Y_j = v_j)$$

$$= \sum_{Z_k} \cdots \sum_{Z_1} P(X_1, \ldots, X_n) Y_1 = v_1, \ldots, Y_j = v_j.$$ 

$$= \sum_{Z_k} \cdots \sum_{Z_1} \prod_{i=1}^n P(X_i | \text{parents}(X_i)) Y_1 = v_1, \ldots, Y_j = v_j.$$
Computation in belief networks reduces to computing the sums of products.

- How can we compute $ab + ac$ efficiently?
Computing sums of products

Computation in belief networks reduces to computing the sums of products.

- How can we compute $ab + ac$ efficiently?
- **Distribute** out the $a$ giving $a(b + c)$
Computing sums of products

Computation in belief networks reduces to computing the sums of products.

- How can we compute $ab + ac$ efficiently?
  - **Distribute** out the $a$ giving $a(b + c)$

- How can we compute $\sum_{Z_1} \prod_{i=1}^{n} P(X_i|\text{parents}(X_i))$ efficiently?
Computing sums of products

Computation in belief networks reduces to computing the sums of products.

- How can we compute \( ab + ac \) efficiently?
- **Distribute** out the \( a \) giving \( a(b + c) \)
- How can we compute \( \sum_{Z_1} \prod_{i=1}^{n} P(X_i|\text{parents}(X_i)) \) efficiently?
- Distribute out those factors that don’t involve \( Z_1 \).
To compute $P(Z|Y_1 = v_1 \land \ldots \land Y_j = v_j)$:

- **Construct a** factor for each conditional probability.
- **Restrict** the observed variables to their observed values.
- **Sum out** each of the other variables (the $\{Z_1, \ldots, Z_k\}$ from slide 45) according to some **elimination ordering**: for each $Z_i$ in order starting from $i = 1$:
  - collect all factors that contain $Z_i$
  - multiply together and sum out $Z_i$
  - add resulting new factor back to the pool
- **Multiply** the remaining factors.
- **Normalize** by dividing the resulting factor $f(Z)$ by $\sum_Z f(Z)$. 
Summing out a variable

To sum out a variable $Z_j$ from a product $f_1, \ldots, f_k$ of factors:

- **Partition** the factors into
  - those that don’t contain $Z_j$, say $f_1, \ldots, f_i$,
  - those that contain $Z_j$, say $f_{i+1}, \ldots, f_k$

We know:

$$\sum_{Z_j} f_1 \times \cdots \times f_k = f_1 \times \cdots \times f_i \times \left( \sum_{Z_j} f_{i+1} \times \cdots \times f_k \right).$$

- **Explicitly construct** a representation of the rightmost factor $\left( \sum_{Z_j} f_{i+1} \times \cdots \times f_k \right)$.
- **Replace** the factors $f_{i+1}, \ldots, f_k$ by the new factor.
Example I

\[
P(A = \text{true}) = 0.3
\]

\[
P(C = \text{true}|A) = \begin{array}{cl} f & 0.8 \\ f & 0.15 \end{array}
\]

\[
P(G = \text{true}|C) = \begin{array}{cl} f & 1.0 \\ f & 0.2 \end{array}
\]

\[
P(L = \text{true}|G) = \begin{array}{cl} f & 0.7 \\ f & 0.2 \end{array}
\]

\[
P(S = \text{true}|L) = \begin{array}{cl} f & 0.9 \\ f & 0.3 \end{array}
\]

see note variableelim.pdf
Notes on VE

- Complexity is **linear** in number of variables, and **exponential** in the size of the largest factor.
- When we create new factors: sometimes this blows up.
- Depends on the **elimination ordering**.
- For **polytrees**: work outside in.
- For general BNs this can be hard.
- Simply finding the optimal elimination ordering is NP-hard for general BNs.
- Inference in general is NP-hard.
Variable Ordering: Polytrees

- eliminate \textbf{singly-connected} nodes \((D, A, C, X_1, \ldots, X_k)\) first
- Then no factor is ever larger than original CPTs
- If you eliminate \(B\) first, a \textbf{large factor} is created that includes \(A, C, X_1, \ldots, X_k\)
Variable Ordering: Relevance

- Certain variables have **no impact**
- In ABC network above, computing $P(A)$ does not require summing over $B$ and $C$

$$P(A) = \sum_{B,C} P(C|B)P(B|A)P(A)$$

$$= P(A) \sum_{B} P(B|A) \sum_{C} P(C|B) = P(A) \times 1.0 \times 1.0$$
Variable Ordering: Relevance

- Can restrict attention to **relevant** variables:
- Given query $Q$ and evidence $E$, **complete** approximation is:
  - $Q$ is relevant
  - if any node is relevant, its parents are relevant
  - if $E \in E$ is a descendant of a relevant variable, then $E$ is relevant
- **irrelevant variable**: a node that is not an ancestor of a query or evidence variable
- this will only remove irrelevant variables, but may not remove them all
Example II

\[ P(M = \text{true}) = 0.08 \]

\[ P(B = \text{true}|M, C) = \]

\[
\begin{array}{c|c|c}
M & C & P(B = \text{true}|M, C) \\
\hline
\text{t} & \text{t} & 0.61 \\
\text{t} & \text{f} & 0.52 \\
\text{f} & \text{t} & 0.78 \\
\text{f} & \text{f} & 0.044 \\
\end{array}
\]

\[ P(C = \text{true}) = 0.32 \]

\[ P(A = \text{true}|C) = \]

\[
\begin{array}{c|c}
C & P(A = \text{true}|C) \\
\hline
\text{t} & 0.80 \\
\text{f} & 0.15 \\
\end{array}
\]

\[ P(D = \text{true}|R) = \]

\[
\begin{array}{c|c}
R & P(D = \text{true}|R) \\
\hline
\text{t} & 0.96 \\
\text{f} & 0.001 \\
\end{array}
\]

\[ P(R = \text{true}|B) = \]

\[
\begin{array}{c|c}
B & P(R = \text{true}|B) \\
\hline
\text{t} & 0.98 \\
\text{f} & 0.01 \\
\end{array}
\]

see note \textit{variableelim.pdf}
Other Representations for Probability distributions

- **Decision Tree** or Graph:

```
    Action
     /   \
    go.out get.coffee
   /     \           /
Rain  Full  /
  t     t 0.1     0.3
0.8  0.6       /   \\
  f     f   \    
```

- **Noisy Or**: \( P(x|Y_1, \ldots, Y_k) \)

- **Logistic Regression**

\[
P(x|Y_1, \ldots, Y_k) = \text{sigmoid}(\sum_i w_i Y_i)
\]

- **Any deep differentiable function** — see A. Stassopoulou and M. Petrou


  https://doi.org/10.1142/S021800149800049X
Reasoning under Uncertainty Part II (Poole & Mackworth (2nd ed.) Chapter 8.5-8.9)