Uncertainty

Why is uncertainty important?
- Agents (and humans) don’t know everything,
- but need to make decisions anyways!
- Decisions are made in the absence of information,
- or in the presence of noisy information (sensor readings)

The best an agent can do:
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 this event
- belief of agent A: based on previous experience of agent A
Uncertainty is epistemological: pertaining to knowledge

Probability: Bayesian

Bayesian probability
all else being equal:
- before 2 flips

Bayesian probability
all else being equal:
- after 2 flips heads, heads

Bayesian probability
all else being equal:
- after 2 flips tails, tails
Probability: Bayesian

Should you wear your seatbelt? estimate \( P(\text{fatality}) \) given you do/don’t wear it

**Frequentist:**
- Sunday (prior to start) - ?
- Monday - 0.0
- Tuesday - 0.0
- Tuesday - 0.33333
- Thursday - 0.25
- Friday - 0.2

... ... ...

N
Number of crashes / N

**Bayesian:**

Features
Describe the world in terms of a set of states: \( \{s_1, s_2, \ldots, s_N\} \)

or, as the product of a set of features (also known as attributes or random variables)

- Number of states = \( 2^{\text{number of binary features}} \)
- Features describe the state space in a factored form.
- state \( \rightarrow \) factorize \( \rightarrow \) feature values
- feature values \( \rightarrow \) cross product \( \rightarrow \) states

Probability Measure

if \( X \) is a random variable (feature, attribute), it can take on values \( x \), where \( x \in \text{Domain}(X) \)

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:

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>0.01</td>
<td>0.7</td>
<td>0.08</td>
<td>0.01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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 \( a \) is definitely false
- \( P(a) = 1 \) means \( a \) 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, \( P(\text{hit a sector}) \)
  - Assume radial independence: \( P(r, \theta) = P(r)P(\theta) \)
  - Only need \( 7 + 3 = 10 \) numbers

Conditional Probability

If \( X \) and \( Y \) are random variables, then
\[ P(x|y) \text{ is the probability that } X = x \text{ given that } Y = y. \]

E.g.
\[ P(\text{flies}|\text{is bird}) \text{ is different than } P(\text{flies}) \]
\[ P(\text{flies}|\text{is a penguin, is bird}) \text{ is different again} \]

Incorporate independence:
\[ P(\text{flies}|\text{is bird, has feathers}) = P(\text{flies}|\text{is bird}) \]

Product rule (Chain rule):
\[ P(\text{flies, is bird}) = P(\text{flies}|\text{is bird})P(\text{is bird}) \]
\[ P(\text{flies, is bird}) = P(\text{is bird}|\text{flies})P(\text{flies}) \]

Leads to: Bayes’ rule
\[ P(\text{is bird}|\text{flies}) = \frac{P(\text{flies}|\text{is bird})P(\text{is bird})}{P(\text{flies})} \]

Sum Rule

We know (an Axiom):
\[ \sum_x P(X = x) = 1.0 \text{ and therefore that } \sum_x P(X = x|Y) = 1.0 \]

This means that
\[ \sum_x P(X = x, Y) = P(Y) \]

Proof:
\[ \sum_x P(X = x, Y) = \sum_x P(X = x|Y)P(Y) \]
\[ = P(Y) \sum_x P(X = x|Y) \]
\[ = P(Y) \]

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):
  \[ E(V) = \sum_{x \in \text{Dom}(X)} P(x)V(x) \]
  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
  \[ E(V(\text{decision})) = \sum_{\text{outcome}} P(\text{outcome}|\text{decision})V(\text{outcome}) \]

Value of Independence

- 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 \( P(X_i|\text{parents}(X_i)) \)

Bayesian networks - example

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

Correlation and Causality

- Directed links in Bayes’ net ≈ causal
- However, not always the case: chocolate → Nobel or Nobel → chocolate?
- In a Bayes net, it doesn’t matter!
- But, some structures will be easier to specify
In this example, it’s probably chocolate ← “Switzerland – ness” → Nobel
Bayesian Networks

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

\[
P(X_i | \text{Parents}(X_i))
\]

for each \(X_i\)

Inference is Easy

Want to know \(P(g)\)? - Use sum rule to sum over parents of \(g\)

\[
P(g) = \sum_{c_i \in \text{Dom}(C)} P(g, c_i)
\]

\[
= \sum_{c_i \in \text{Dom}(C)} P(g | c_i) P(c_i)
\]

\[
= \sum_{c_i \in \text{Dom}(C)} P(g | c_i) \sum_{a_j \in \text{Dom}(A)} P(c_i | a_j) P(a_j)
\]

\[
= P(g | c) [P(c | a) P(a) + P(c | \bar{a}) P(\bar{a})]
\]

\[
+ P(g | \bar{c}) [P(\bar{c} | a) P(a) + P(\bar{c} | \bar{a}) P(\bar{a})]
\]

\[
= 1.0 \times [0.8 \times 0.3 + 0.15 \times 0.7] + 0.2 \times [0.2 \times 0.3 + 0.85 \times 0.7]
\]

\[
= 0.4760
\]
The structure of the BN means that:

\[ P(X_i | S, \text{Parents}(X_i)) = P(X_i | \text{Parents}(X_i)) \]

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

The BN defines a **factorization** of the joint probability distribution. The joint distribution is formed by multiplying the conditional probability tables together.

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**

**Three Basic Bayesian Networks**

- **Test B** and **Test A** are independent if **Cancer** is observed
- **Test B** and **Test A** are independent if **Database** is observed
Malfunction and Cancer are independent if Test B is not observed.

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?

The Markov Blanket of a node 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 of 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)?

Why is Bayes' theorem interesting?

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

Probabilistic Inference

- Agent has a prior belief in a hypothesis, \( h, P(h) \),
- Agent observes some evidence \( e \) that has a likelihood given the hypothesis: \( P(e \mid h) \).
- The agent's posterior belief about \( h \) after observing \( e \), \( P(h \mid e) \), is given by Bayes' Rule:
  \[
P(h \mid e) = \frac{P(e \mid h)P(h)}{P(e)} = \frac{P(e \mid h)P(h)}{\sum_h P(e \mid h)P(h)}
  \]

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 so the evidence gives Database=true (not directly Test B= true) we want to know \( P(\text{Cancer} = \text{true} \mid \text{Database} = \text{true}) \)
- \( P(\text{Cancer} = \text{true} \mid \text{Database} = \text{true}) = 0.80 \)
- \( P(\text{Malfunction} = \text{true} \mid \text{Database} = \text{true}) = 0.14 \)
  (we will see how to get these numbers later)

Suppose Test A is positive as well we want \( P(\text{Cancer} = \text{true} \mid \text{Database} = \text{true} \land \text{Test A} = \text{true}) \)
- \( P(\text{Cancer} = \text{true} \mid \text{Database} = \text{true} \land \text{Test A} = \text{true}) = 0.95 \)
- \( P(\text{Malfunction} = \text{true} \mid \text{Database} = \text{true} \land \text{Test A} = \text{true}) = 0.08 \)
  (we will see how to get these numbers later)
Probabilistic Inference

Suppose Test A is negative, though! we want \( P(\text{Cancer} = \text{true} | \text{Database} = \text{true} \land \text{TestA} = \text{false}) \)
- \( P(\text{Cancer} = \text{true} | \text{Database} = \text{true} \land \text{TestA} = \text{false}) = 0.48 \)
- \( P(M = \text{true} | \text{Database} = \text{true} \land \text{TestA} = \text{false}) = 0.27 \)
(we will see how to get these numbers later)

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 Backward Inference

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

\[
P(ET | j) \propto P(j, ET)
= \sum_M P(j | M) P(M | ET) P(ET)
= \sum_M P(j | M) P(M | ET) P(ET)
\]
normalising constant is \( \frac{1}{P(j)} \), but this can be computed as

\[
P(j) = \sum_{ET} P(ET, j)
\]
Backward Inference

Backward Inference

http://imgs.xkcd.com/comics/bridge.png

F: Bridge on Fire
C: All friends Crazy
J: All friends Jump
What is $P(F|J = \text{true})$?

Factors

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:

- $f(X_1 = v_1, X_2, \ldots, X_j)$, 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.

Multiplying factors

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:

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

Example factors

Multiplying factors example

<table>
<thead>
<tr>
<th>$X$</th>
<th>$Y$</th>
<th>$Z$</th>
<th>val</th>
</tr>
</thead>
<tbody>
<tr>
<td>t</td>
<td>t</td>
<td>t</td>
<td>0.1</td>
</tr>
<tr>
<td>t</td>
<td>t</td>
<td>f</td>
<td>0.9</td>
</tr>
<tr>
<td>t</td>
<td>f</td>
<td>t</td>
<td>0.2</td>
</tr>
<tr>
<td>f</td>
<td>t</td>
<td>t</td>
<td>0.4</td>
</tr>
<tr>
<td>f</td>
<td>t</td>
<td>f</td>
<td>0.6</td>
</tr>
<tr>
<td>f</td>
<td>f</td>
<td>t</td>
<td>0.3</td>
</tr>
<tr>
<td>f</td>
<td>f</td>
<td>f</td>
<td>0.7</td>
</tr>
</tbody>
</table>

Multiplying factors example

<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>

<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>

Example factors

<table>
<thead>
<tr>
<th>$X$</th>
<th>$Y$</th>
<th>$Z$</th>
<th>val</th>
</tr>
</thead>
<tbody>
<tr>
<td>t</td>
<td>t</td>
<td>t</td>
<td>0.1</td>
</tr>
<tr>
<td>t</td>
<td>t</td>
<td>f</td>
<td>0.9</td>
</tr>
<tr>
<td>t</td>
<td>f</td>
<td>t</td>
<td>0.2</td>
</tr>
<tr>
<td>f</td>
<td>t</td>
<td>t</td>
<td>0.4</td>
</tr>
<tr>
<td>f</td>
<td>t</td>
<td>f</td>
<td>0.8</td>
</tr>
<tr>
<td>f</td>
<td>f</td>
<td>t</td>
<td>0.3</td>
</tr>
<tr>
<td>f</td>
<td>f</td>
<td>f</td>
<td>0.7</td>
</tr>
</tbody>
</table>

Variable Elimination

- intuitions above: polytree algorithm
- works for simple networks without loops
- more general algorithm: Variable Elimination
- applies sum-out rule repeatedly
- distributes sums

Example factors

Multiplying factors example

<table>
<thead>
<tr>
<th>$X$</th>
<th>$Y$</th>
<th>$Z$</th>
<th>val</th>
</tr>
</thead>
<tbody>
<tr>
<td>t</td>
<td>t</td>
<td>t</td>
<td>0.1</td>
</tr>
<tr>
<td>t</td>
<td>t</td>
<td>f</td>
<td>0.9</td>
</tr>
<tr>
<td>t</td>
<td>f</td>
<td>t</td>
<td>0.2</td>
</tr>
<tr>
<td>f</td>
<td>t</td>
<td>t</td>
<td>0.4</td>
</tr>
<tr>
<td>f</td>
<td>t</td>
<td>f</td>
<td>0.8</td>
</tr>
<tr>
<td>f</td>
<td>f</td>
<td>t</td>
<td>0.3</td>
</tr>
<tr>
<td>f</td>
<td>f</td>
<td>f</td>
<td>0.7</td>
</tr>
</tbody>
</table>

Example factors

Multiplying factors example

<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>

Example factors

Multiplying factors example

<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>

Example factors

Multiplying factors example

<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>0.1</td>
</tr>
<tr>
<td>t</td>
<td>t</td>
<td>f</td>
<td>0.9</td>
</tr>
<tr>
<td>t</td>
<td>f</td>
<td>t</td>
<td>0.2</td>
</tr>
<tr>
<td>f</td>
<td>t</td>
<td>t</td>
<td>0.4</td>
</tr>
<tr>
<td>f</td>
<td>t</td>
<td>f</td>
<td>0.8</td>
</tr>
<tr>
<td>f</td>
<td>f</td>
<td>t</td>
<td>0.3</td>
</tr>
<tr>
<td>f</td>
<td>f</td>
<td>f</td>
<td>0.7</td>
</tr>
</tbody>
</table>
Summing out variables

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:

\[
\left(\sum_{x_1} f(x_2, \ldots, x_j)\right)
= f(X_1 = v_1, \ldots, X_j) + \cdots + f(X_1 = v_k, \ldots, X_j)
\]

Evidence

If we want to compute the posterior probability of $Z$ given evidence $Y_1 = v_1 \land \ldots \land Y_j = v_j$:

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

So the computation reduces to the probability of $P(Z, Y_1 = v_1, \ldots, Y_j = v_j)$. We 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 other variables, $Z_1, \ldots, Z_k = \{X_1, \ldots, X_n\} - \{Z\} - \{Y_1, \ldots, Y_j\}$. We order the $Z_i$ into an **elimination ordering**.

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

Computing sums of products

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?

Variable elimination algorithm

To compute \( P(Z|Y_1 = v_1 \land \ldots \land Y_j = v_j) \):

- Construct a factor for each conditional probability.
- Set the observed variables to their observed values.
- Sum out each of the other variables (the \( \{Z_1, \ldots, Z_k\} \)) according to some elimination ordering.
- 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. Replace the factors \( f_{i+1}, \ldots, f_k \) by the new factor.

Example I

A C G L S

\( P(A = \text{true}) = 0.3 \) \( P(C = \text{true}|A) = A \) \( P(C = \text{true}|A) = 0.8 \) \( P(G = \text{true}|C) = C \) \( P(G = \text{true}|C) = 1.0 \) \( P(L = \text{true}|G) = G \) \( P(L = \text{true}|G) = 0.7 \) \( P(S = \text{true}|L) = L \) \( P(S = \text{true}|L) = 0.9 \)

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 "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 factor is created that includes \(A, C, X_1, \ldots, X_k\)
- e.g.

\[
P(D|x_1) = \sum_{A,C} P(D|A)P(A)P(C) \sum_{B,X_2,\ldots,X_k} P(B|A,C)P(x_1|B) \prod_{j=2}^k P(X_j|B)
\]

Example II

Other Representations for Probability distributions
- Decision Tree:
- Noisy Or: \(P(x|Y_1, \ldots, Y_k)\) where \(X\) is caused by parents \(Y\)
- Logistic Regression where parents \(Y\) caused by \(X\).

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