Reasoning under Uncertainty

- Uncertainty
- Probability
- Syntax
- Semantics
- Inference rules

A little history…

- Early AI researchers largely rejected using probability in their systems
  - "People don't think that way…"

- But symbolic approaches don't tolerate approximation well…

Uncertainty

Let action A_t = leave for airport t minutes before flight
Will A_t get me there on time?

Problems:

1) partial observability (road state, other drivers' plans, etc.)
2) noisy sensors (KCBS traffic reports)
3) uncertainty in action outcomes (flat tire, etc.)
4) immense complexity of modelling and predicting traffic

Hence a purely logical approach either

1) risks falsehood: "A_{25} will get me there on time"

or 2) leads to conclusions that are too weak for decision making:
"A_{25} will get me there on time if there's no accident on the bridge and it doesn't rain and my tires remain intact etc etc."

(A_{1440} might reasonably be said to get me there on time
but I'd have to stay overnight in the airport ...)

Methods for handling uncertainty

Default or nonmonotonic logic:

Assume my car does not have a flat tire

Assume A_{25} works unless contradicted by evidence

Issues: What assumptions are reasonable? How to handle contradiction?

Rules with fudge factors:

$A_{25} \mapsto_{0.3}$ get there on time

$Sprinkler \mapsto_{0.99} WetGrass$

$WetGrass \mapsto_{0.7} Rain$

Issues: Problems with combination, e.g., $Sprinkler$ causes $Rain$??

Probability

Given the available evidence,

A_{25} will get me there on time with probability 0.04

Mahaviracarya (9th C.), Cardamo (1565) theory of gambling

(Fuzzy logic handles *degree of truth* NOT uncertainty e.g.,

$WetGrass$ is true to degree 0.2)

Probability

Probabilistic assertions *summarize* effects of
laziness: failure to enumerate exceptions, qualifications, etc.
ignorance: lack of relevant facts, initial conditions, etc.

Subjective or Bayesian probability:
Probabilities relate propositions to one's own state of knowledge
e.g., $P(A_{25}|\text{no reported accidents}) = 0.06$

These are not assertions about the world

Probabilities of propositions change with new evidence:
e.g., $P(A_{25}|\text{no reported accidents, 5 a.m.}) = 0.15$

(Analogous to logical entailment status $KB \models \alpha$, not truth.)

Making decisions under uncertainty

Suppose I believe the following:

$$P(A_{25} \text{ gets me there on time} | \ldots) = 0.04$$
$$P(A_{90} \text{ gets me there on time} | \ldots) = 0.70$$
$$P(A_{120} \text{ gets me there on time} | \ldots) = 0.95$$
$$P(A_{1440} \text{ gets me there on time} | \ldots) = 0.9999$$

Which action to choose?

Depends on my preferences for missing flight vs. airport cuisine, etc.

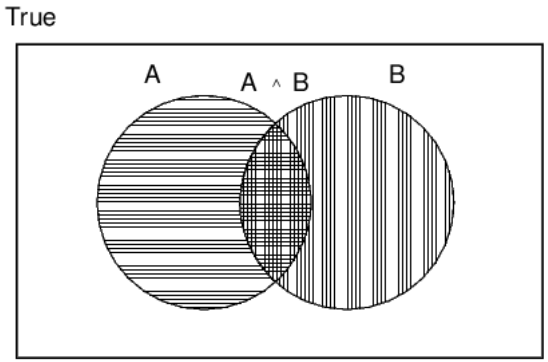
Utility theory is used to represent and infer preferences

Decision theory = utility theory + probability theory

Axioms of probability

For any propositions A, B

1. $0 \leq P(A) \leq 1$
2. $P(True) = 1$ and $P(False) = 0$
3. $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$



de Finetti (1931): an agent who bets according to probabilities that violate these axioms can be forced to bet so as to lose money regardless of outcome.

Probability basics

Begin with a set Ω—the sample space

e.g., 6 possible rolls of a die.

$\omega \in \Omega$ is a sample point/possible world/atomic event

A probability space or probability model is a sample space with an assignment $P(\omega)$ for every $\omega \in \Omega$ s.t.

$0 \leq P(\omega) \leq 1$

$\sum_\omega P(\omega) = 1$

e.g., $P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = 1/6$.

An event A is any subset of Ω

$$P(A) = \sum_{\{\omega \in A\}} P(\omega)$$

E.g., $P(\text{die roll} < 4) = P(1) + P(2) + P(3) = 1/6 + 1/6 + 1/6 = 1/2$

Random variables

A random variable is a function from sample points to some range, e.g., the reals or Booleans

e.g., $Odd(1) = true$.

P induces a probability distribution for any r.v. X:

$$P(X = x_i) = \Sigma_{\{\omega : X(\omega) = x_i\}} P(\omega)$$

e.g., $P(Odd = true) = P(1) + P(3) + P(5) = 1/6 + 1/6 + 1/6 = 1/2$

Propositions

Think of a proposition as the event (set of sample points)
where the proposition is true

Given Boolean random variables A and B:

event a = set of sample points where $A(\omega) = true$
event $\neg a$ = set of sample points where $A(\omega) = false$
event $a \wedge b$ = points where $A(\omega) = true$ and $B(\omega) = true$

Often in AI applications, the sample points are **defined**
by the values of a set of random variables, i.e., the
sample space is the Cartesian product of the ranges of the variables

With Boolean variables, sample point = propositional logic model

e.g., $A = true$, $B = false$, or $a \wedge \neg b$.

Proposition = disjunction of atomic events in which it is true

e.g., $(a \vee b) \equiv (\neg a \wedge b) \vee (a \wedge \neg b) \vee (a \wedge b)$

$\Rightarrow P(a \vee b) = P(\neg a \wedge b) + P(a \wedge \neg b) + P(a \wedge b)$

Syntax

Similar to propositional logic: possible worlds defined by assignment of values to random variables.

Propositional or Boolean random variables
e.g., $Cavity$ (do I have a cavity?)
Include propositional logic expressions
e.g., $\neg Burglary \vee Earthquake$

Multivalued random variables
e.g., $Weather$ is one of $\langle sunny, rain, cloudy, snow \rangle$
Values must be exhaustive and mutually exclusive

Proposition constructed by assignment of a value:
e.g., $Weather = sunny$; also $Cavity = true$ for clarity

Syntax

Prior or unconditional probabilities of propositions
e.g., $P(Cavity) = 0.1$ and $P(Weather = sunny) = 0.72$
correspond to belief prior to arrival of any (new) evidence

Probability distribution gives values for all possible assignments:
$\mathbf{P}(Weather) = \langle 0.72, 0.1, 0.08, 0.1 \rangle$ (normalized, i.e., sums to 1)

Joint probability distribution for a set of variables gives values for each possible assignment to all the variables
$\mathbf{P}(Weather, Cavity)$ = a 4×2 matrix of values:

$Weather =$	$sunny$	$rain$	$cloudy$	$snow$
$Cavity = true$	0.144	0.02	0.016	0.02
$Cavity = false$	0.576	0.08	0.064	0.08

Sum of all entries is 1

Syntax

Conditional or posterior probabilities

e.g., $P(Cavity|Toothache) = 0.8$

i.e., given that $Toothache$ is all I know

Notation for conditional distributions:

$\mathbf{P}(Weather|Earthquake) =$ Gives probabilities for all values of *Weather* and *Earthquake*

If we know more, e.g., $Cavity$ is also given, then we have

$P(Cavity|Toothache, Cavity) = 1$

Note: the less specific belief $remains\ valid$ after more evidence arrives, but is not always $useful$

New evidence may be irrelevant, allowing simplification, e.g.,

$P(Cavity|Toothache, 49ersWin) = P(Cavity|Toothache) = 0.8$

This kind of inference, sanctioned by domain knowledge, is crucial

Conditional probability

Definition of conditional probability:

$$P(A|B) = \frac{P(A \wedge B)}{P(B)} \text{ if } P(B) \neq 0$$

Product rule gives an alternative formulation:

$$P(A \wedge B) = P(A|B)P(B) = P(B|A)P(A)$$

A general version holds for whole distributions, e.g.,

$$\mathbf{P}(Weather, Cavity) = \mathbf{P}(Weather|Cavity)\mathbf{P}(Cavity)$$

(View as a 4×2 set of equations, *not* matrix mult.)

Chain rule is derived by successive application of product rule:

$$\begin{aligned}\mathbf{P}(X_1, \ldots, X_n) &= \mathbf{P}(X_1, \ldots, X_{n-1})\ \mathbf{P}(X_n|X_1, \ldots, X_{n-1}) \\ &= \mathbf{P}(X_1, \ldots, X_{n-2})\ \mathbf{P}(X_{n_1}|X_1, \ldots, X_{n-2})\ \mathbf{P}(X_n|X_1, \ldots, X_{n-1}) \\ &= \ldots \\ &= \prod_{i=1}^{n} \mathbf{P}(X_i|X_1, \ldots, X_{i-1})\end{aligned}$$

Bayes' rule

Product rule $P(A \wedge B) = P(A|B)P(B) = P(B|A)P(A)$

$$\Rightarrow \underline{\text{Bayes' rule}}\ P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Why is this useful???

For assessing diagnostic probability from causal probability:

$$P(Cause|Effect) = \frac{P(Effect|Cause)P(Cause)}{P(Effect)}$$

E.g., let M be meningitis, S be stiff neck:

$$P(M|S) = \frac{P(S|M)P(M)}{P(S)} = \frac{0.8 \times 0.0001}{0.1} = 0.0008$$

Note: posterior probability of meningitis still very small!

Normalization

Suppose we wish to compute a posterior distribution over A given $B = b$, and suppose A has possible values $a_1 \ldots a_m$

We can apply Bayes' rule for each value of A:

$$P(A = a_1 | B = b) = P(B = b | A = a_1) P(A = a_1) / P(B = b)$$

$$\ldots$$

$$P(A = a_m | B = b) = P(B = b | A = a_m) P(A = a_m) / P(B = b)$$

Adding these up, and noting that $\Sigma_i P(A = a_i | B = b) = 1$:

$$1 / P(B = b) = 1 / \Sigma_i P(B = b | A = a_i) P(A = a_i)$$

This is the normalization factor, constant w.r.t. i, denoted α:

$$\mathbf{P}(A | B = b) = \alpha \mathbf{P}(B = b | A) \mathbf{P}(A)$$

i.e., we do not need to spend effort trying to know P(B)

Typically compute an unnormalized distribution, normalize at end

e.g., suppose $\mathbf{P}(B = b | A) \mathbf{P}(A) = \langle 0.4, 0.2, 0.2 \rangle$

then $\mathbf{P}(A | B = b) = \alpha \langle 0.4, 0.2, 0.2 \rangle = \frac{\langle 0.4, 0.2, 0.2 \rangle}{0.4 + 0.2 + 0.2} = \langle 0.5, 0.25, 0.25 \rangle$

Conditioning

Introducing a variable as an extra condition:

$$P(X|Y) = \Sigma_z P(X|Y, Z=z)P(Z=z|Y)$$

Intuition: often easier to assess each specific circumstance, e.g., $P(RunOver|Cross)$

$$\begin{aligned} &= P(RunOver|Cross, Light=green)P(Light=green|Cross) \\ &+ P(RunOver|Cross, Light=yellow)P(Light=yellow|Cross) \\ &+ P(RunOver|Cross, Light=red)P(Light=red|Cross) \end{aligned}$$

When Y is absent, we have summing out or marginalization:

$$P(X) = \Sigma_z P(X|Z=z)P(Z=z) = \Sigma_z P(X, Z=z)$$

In general, given a joint distribution over a set of variables, the distribution over any subset (called a marginal distribution for historical reasons) can be calculated by summing out the other variables.

Full joint distributions

A complete probability model specifies every entry in the joint distribution for all the variables $\mathbf{X} = X_1, \ldots, X_n$
I.e., a probability for each possible world $X_1 = x_1, \ldots, X_n = x_n$

(Cf. complete theories in logic.)

E.g., suppose $Toothache$ and $Cavity$ are the random variables:

	$Toothache = true$	$Toothache = false$
$Cavity = true$	0.04	0.06
$Cavity = false$	0.01	0.89

Possible worlds are mutually exclusive $\Rightarrow$ $P(w_1 \wedge w_2) = 0$
Possible worlds are exhaustive $\Rightarrow$ $w_1 \vee \cdots \vee w_n$ is $True$
hence $\sum_i P(w_i) = 1$

Full joint distribution

1) For any proposition ϕ defined on the random variables
$\phi(w_i)$ is true or false

2) ϕ is equivalent to the disjunction of w_is where $\phi(w_i)$ is true

Hence $P(\phi) = \Sigma_{\{w_i:\ \phi(w_i)\}} P(w_i)$

I.e., the unconditional probability of any proposition is computable as the sum of entries from the full joint distribution

Conditional probabilities can be computed in the same way as a ratio:

$$P(\phi|\xi) = \frac{P(\phi \wedge \xi)}{P(\xi)}$$

E.g.,

$$P(Cavity|Toothache) = \frac{P(Cavity \wedge Toothache)}{P(Toothache)} = \frac{0.04}{0.04 + 0.01} = 0.8$$

Inference by enumeration

Start with the joint distribution:

	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	**.108**	**.012**	**.072**	**.008**
¬ *cavity*	**.016**	**.064**	**.144**	**.576**

For any proposition ϕ, sum the atomic events where it is true:

$P(\phi) = \sum_{\omega:\omega \models \phi} P(\omega)$

Inference by enumeration

Start with the joint distribution:

	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	**.108**	**.012**	**.072**	**.008**
¬ *cavity*	**.016**	**.064**	**.144**	**.576**

For any proposition ϕ, sum the atomic events where it is true:

$P(\phi) = \sum_{\omega:\omega \models \phi} P(\omega)$

$P(toothache) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2$

Inference by enumeration

Start with the joint distribution:

	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	**.108**	**.012**	**.072**	**.008**
¬ *cavity*	**.016**	**.064**	**.144**	**.576**

For any proposition ϕ, sum the atomic events where it is true:

$P(\phi) = \sum_{\omega:\omega \models \phi} P(\omega)$

$P(cavity \vee toothache) = 0.108 + 0.012 + 0.072 + 0.008 + 0.016 + 0.064 = 0.28$

Inference by enumeration

Start with the joint distribution:

	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	.108	.012	.072	.008
¬ *cavity*	.016	.064	.144	.576

Can also compute conditional probabilities:

$$P(\neg cavity|toothache) = \frac{P(\neg cavity \wedge toothache)}{P(toothache)}$$
$$= \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064} = 0.4$$

Inference by enumeration

	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	.108	.012	.072	.008
¬ *cavity*	.016	.064	.144	.576

Uppercase name:
For the whole distribution

Denominator can be viewed as a normalization constant α

$$\begin{aligned}\mathbf{P}(Cavity|toothache) &= \alpha\,\mathbf{P}(Cavity, toothache)\\ &= \alpha\,[\mathbf{P}(Cavity, toothache, catch) + \mathbf{P}(Cavity, toothache, \neg catch)]\\ &= \alpha\,[\langle 0.108, 0.016\rangle + \langle 0.012, 0.064\rangle]\\ &= \alpha\,\langle 0.12, 0.08\rangle = \langle 0.6, 0.4\rangle\end{aligned}$$

General idea: compute distribution on query variable
by fixing evidence variables and summing over hidden variables

Inference from joint distributions

Typically, we are interested in
the posterior joint distribution of the query variables $\mathbf{Y}$
given specific values $\mathbf{e}$ for the evidence variables $\mathbf{E}$

Let the hidden variables be $\mathbf{H} = \mathbf{Y} - \mathbf{E}$

Then the required summation of joint entries is done by summing out the hidden variables:

$$\mathbf{P}(\mathbf{Y}|\mathbf{E}=\mathbf{e}) = \alpha\mathbf{P}(\mathbf{Y},\mathbf{E}=\mathbf{e}) = \alpha\Sigma_{\mathbf{h}}\mathbf{P}(\mathbf{Y},\mathbf{E}=\mathbf{e},\mathbf{H}=\mathbf{h})$$

The terms in the summation are joint entries because $\mathbf{Y}$, $\mathbf{E}$, and $\mathbf{H}$ together exhaust the set of random variables

Obvious problems:

1) Worst-case time complexity $O(d^n)$ where d is the largest arity
2) Space complexity $O(d^n)$ to store the joint distribution
3) How to find the numbers for $O(d^n)$ entries???

Example

Assume the full joint distribution:

	toothache		¬toothache	
	catch	¬catch	catch	¬catch
cavity	0.108	0.012	0.072	0.008
¬cavity	0.016	0.064	0.144	0.576

Y=Cavity: we want to know whether we have a cavity

E=Toothache: we know we have a toothache

H=Catch: we don't know whether probe would catch or not

$$P(Y|e) = \alpha\ P(Y, e) = \alpha \sum_h P(Y, e, h) = \frac{0.108 + 0.012}{0.108+0.012+0.016+0.064}$$

Belief networks

- Conditional independence
- Syntax and semantics
- Exact inference
- Approximate inference

Independence

Two random variables A B are (absolutely) independent iff

$$P(A|B) = P(A)$$

or $P(A, B) = P(A|B)P(B) = P(A)P(B)$

e.g., A and B are two coin tosses

If n Boolean variables are independent, the full joint is

$$\mathbf{P}(X_1, \ldots, X_n) = \prod_i \mathbf{P}(X_i)$$

hence can be specified by just n numbers

Absolute independence is a very strong requirement, seldom met

Conditional independence

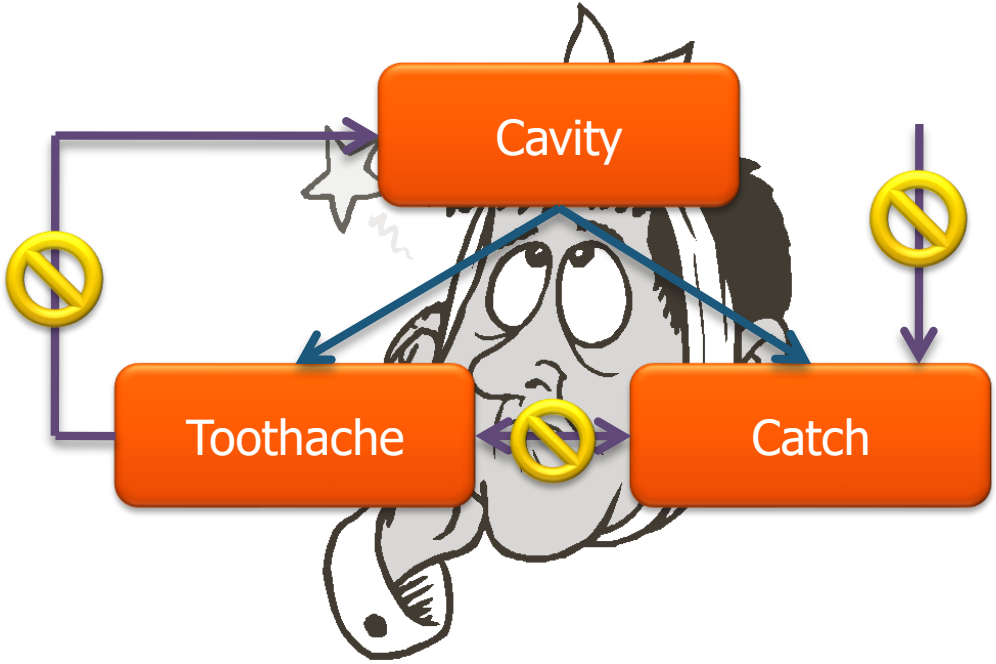
Consider the dentist problem with three random variables:
$Toothache$, $Cavity$, $Catch$ (steel probe catches in my tooth)

The full joint distribution has $2^3 - 1 = 7$ independent entries

If I have a cavity, the probability that the probe catches in it doesn't depend on whether I have a toothache:
(1) $P(Catch|Toothache, Cavity) = P(Catch|Cavity)$
i.e., $Catch$ is conditionally independent of $Toothache$ given $Cavity$

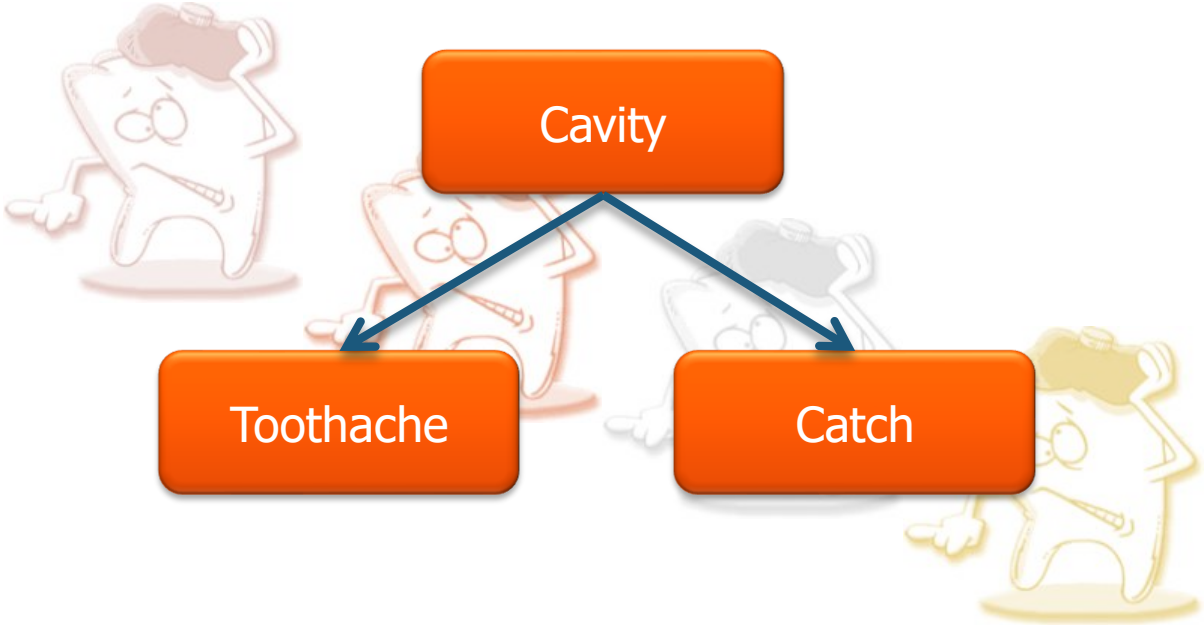
The same independence holds if I haven't got a cavity:
(2) $P(Catch|Toothache, \neg Cavity) = P(Catch|\neg Cavity)$

Conditional Independence



Other interactions may exist, but they are either insignificant, unknown or irrelevant. We leave them out.

Conditional Independence



We **assume** that a "catch" is not influenced by a toothache and visa versa.

Conditional independence

Equivalent statements to (1)

(1a) $P(Toothache|Catch, Cavity) = P(Toothache|Cavity)$ Why??

(1b) $P(Toothache, Catch|Cavity) = P(Toothache|Cavity)P(Catch|Cavity)$
Why??

Full joint distribution can now be written as

$$\mathbf{P}(Toothache, Catch, Cavity) = \mathbf{P}(Toothache, Catch|Cavity)\mathbf{P}(Cavity)$$
$$= \mathbf{P}(Toothache|Cavity)\mathbf{P}(Catch|Cavity)\mathbf{P}(Cavity)$$

i.e., $2 + 2 + 1 = 5$ independent numbers (equations 1 and 2 remove 2)

Conditional independence

Equivalent statements to (1)

(1a) $P(Toothache|Catch, Cavity) = P(Toothache|Cavity)$ Why??

$$
\begin{aligned}
&P(Toothache|Catch, Cavity) \\
&\quad = P(Catch|Toothache, Cavity)P(Toothache|Cavity)/P(Catch|Cavity) \quad \text{(Bayes)} \\
&\quad = P(Catch|Cavity)P(Toothache|Cavity)/P(Catch|Cavity) \text{ (from 1)} \\
&\quad = P(Toothache|Cavity)
\end{aligned}
$$

(1b) $P(Toothache, Catch|Cavity) = P(Toothache|Cavity)P(Catch|Cavity)$
Why??

$$
\begin{aligned}
&P(Toothache, Catch|Cavity) \\
&\quad = P(Toothache|Catch, Cavity)P(Catch|Cavity) \text{ (product rule)} \\
&\quad = P(Toothache|Cavity)P(Catch|Cavity) \text{ (from 1a)}
\end{aligned}
$$

Belief networks

A simple, graphical notation for conditional independence assertions and hence for compact specification of full joint distributions

Syntax:

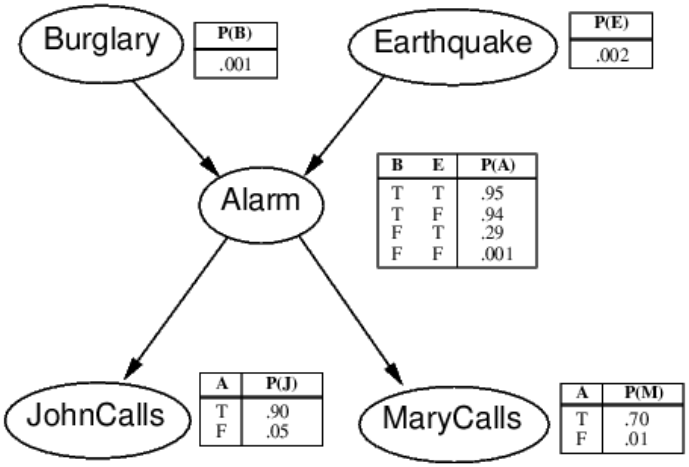
- a set of nodes, one per variable
- a directed, acyclic graph (link $\approx$ "directly influences")
- a conditional distribution for each node given its parents:
  $\mathbf{P}(X_i|Parents(X_i))$

In the simplest case, conditional distribution represented as a conditional probability table (CPT)

Example

I'm at work, neighbor John calls to say my alarm is ringing, but neighbor Mary doesn't call. Sometimes it's set off by minor earthquakes. Is there a burglar?

Variables: $Burglar$, $Earthquake$, $Alarm$, $JohnCalls$, $MaryCalls$
Network topology reflects "causal" knowledge:



Note: $\leq k$ parents $\Rightarrow$ $O(d^k n)$ numbers vs. $O(d^n)$

Semantics

"Global" semantics defines the full joint distribution as the product of the local conditional distributions:

$$\mathbf{P}(X_1, \ldots, X_n) = \prod_{i=1}^{n} \mathbf{P}(X_i | Parents(X_i))$$

e.g., $P(J \wedge M \wedge A \wedge \neg B \wedge \neg E)$ is given by??

$=$

Semantics

"Global" semantics defines the full joint distribution as the product of the local conditional distributions:

$$\mathbf{P}(X_1, \ldots, X_n) = \prod_{i=1}^{n} \mathbf{P}(X_i | Parents(X_i))$$

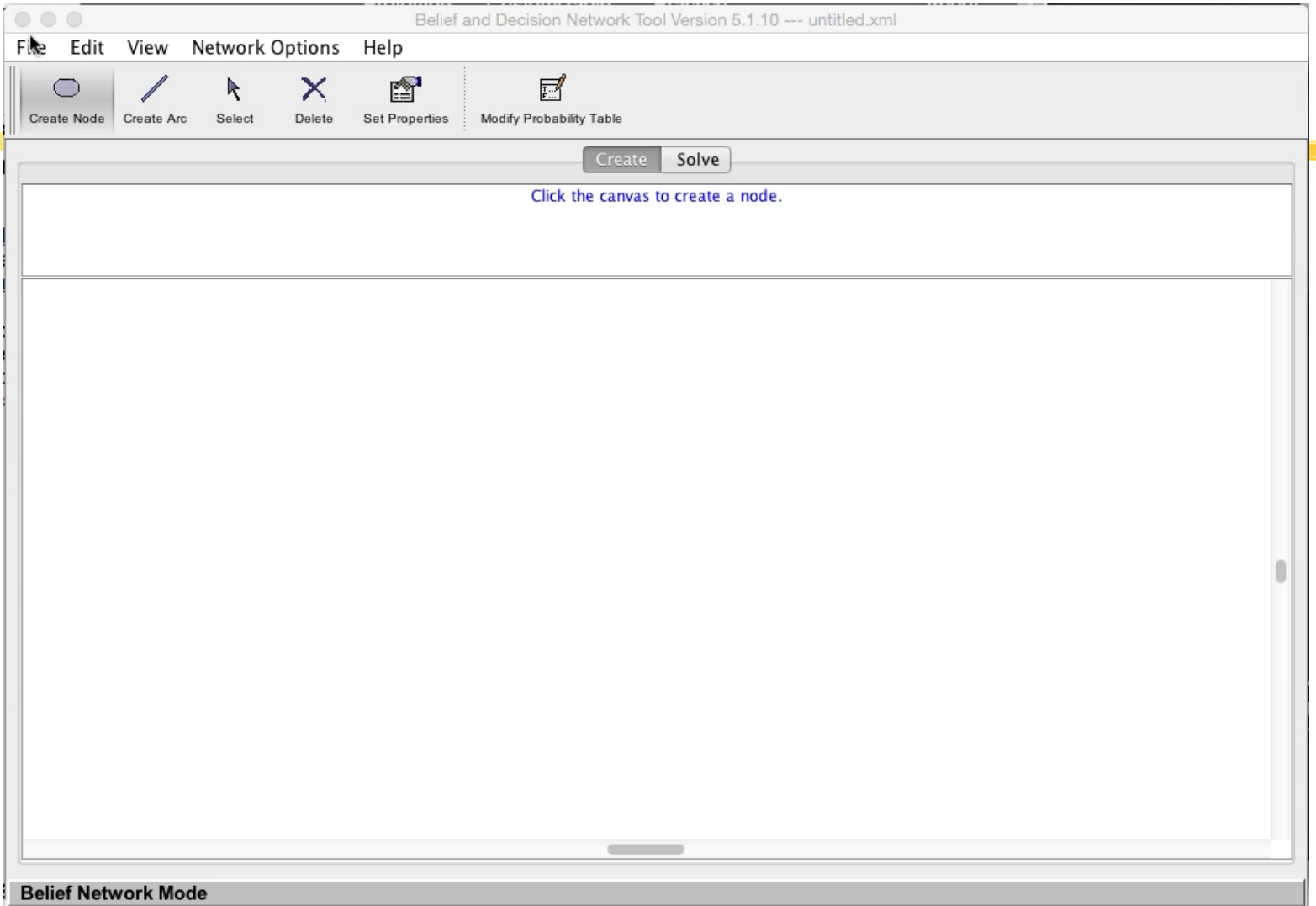
e.g., $P(J \wedge M \wedge A \wedge \neg B \wedge \neg E)$ is given by??

$= P(\neg B)P(\neg E)P(A|\neg B \wedge \neg E)P(J|A)P(M|A)$

"Local" semantics: each node is conditionally independent of its nondescendants given its parents

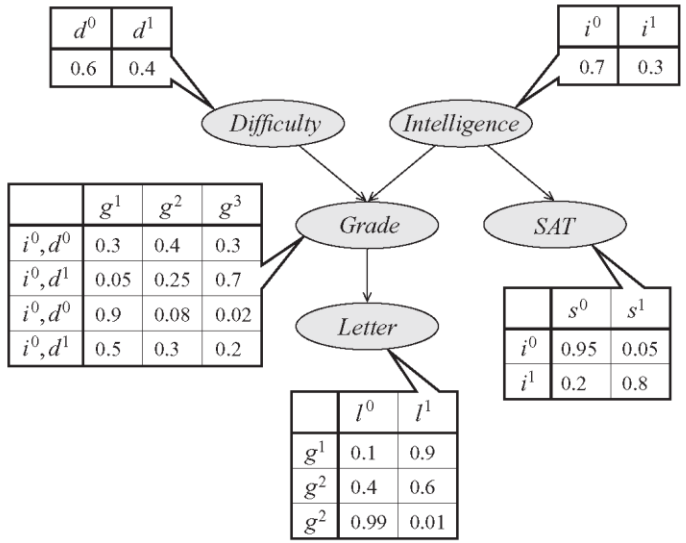
Theorem: Local semantics $\Leftrightarrow$ global semantics

Demo


Belief and Decision Network Tool Version 5.1.10 --- untitled.xml
File
Edit
View
Network Options
Help
Create Node
Create Arc
Select
Delete
Set Properties
Modify Probability Table
Create
Solve
Click the canvas to create a node.
Belief Network Mode

Bayesian Network: Student Model

Graph and CPDs



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^0, d^0	0.9	0.08	0.02
i^0, d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^2	0.99	0.01

typo: these two { rows are for i^1

$Val(I)=\{i^0=$low intelligence, $i^1=$ high intelligence$\}$

$Val(D)=\{d^0=$easy, $d^1=$hard$\}$

$Val(G)=\{g^1=A,\ g^2=B,\ g^3=C\}$

$Val(S)=\{s^0=$low, $s^1=$high$)$

$Val(L)=\{l^0=$weak, $l^1=$strong$\}$

$$P(D,I,G,S,L) = P(D)P(I)P(G \mid D,I)P(S \mid I)P(L \mid G)$$

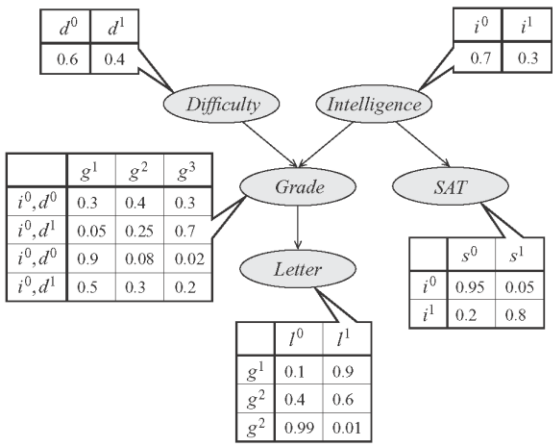
$$P(i^1,d^0,g^2,s^1,l^0) = P(i^1)P(d^0)P(g^2 \mid i^1,d^0)P(s^1 \mid i^1)P(l^0 \mid g^2)$$

$$=0.3 \cdot 0.6 \cdot 0.08 \cdot 0.8 \cdot 0.4 = 0.004608$$

Chain rule for Bayesian network

Reasoning Patterns

Reasoning about a student George using the model



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^0, d^0	0.9	0.08	0.02
i^0, d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^2	0.99	0.01

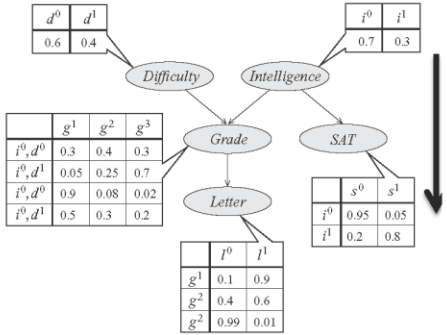
- Causal Reasoning
  - George is interested in knowing as to how likely he is to get a strong letter (based on intelligence, difficulty)?
- Evidential Reasoning
  - Recruiter is interested in knowing whether George is intelligent (based on letter, SAT)

Causal Reasoning

1. How likely is George to get a strong letter (knowing nothing else)?
   - $P(l^1)=0.502$
   - Obtained by summing-out other variables in joint distribution
2. But George is not so intelligent (i^0)
   - $P(l^1|i^0)=0.389$
3. Next we find out ECON101 is easy (d^0)
   - $P(l^1|i^0, d^0)=0.513$

Observe how probabilities change as evidence is obtained

$$P(D,I,G,S,l^1) = \sum_{D,I,G,S} P(D)P(I)P(G|D,I)P(S|I)P(l^1|G)$$

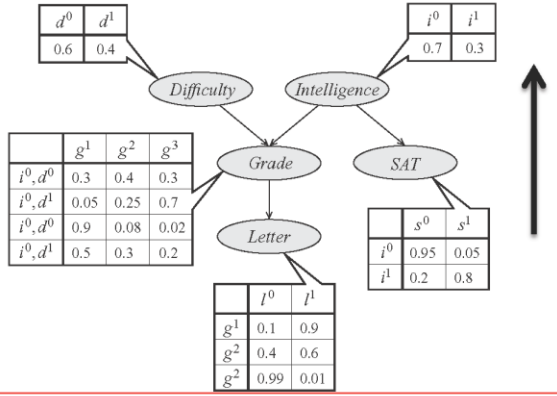


Query is Example of Causal Reasoning:

Predicting downstream effects of factors such as intelligence

Evidential Reasoning

- Recruiter wants to hire intelligent student
- A priori George is 30% likely to be intelligent
- $P(i^1)=0.3$
- Finds that George received grade C (g^3) in ECON101
- $P(i^1|g^3)=0.079$
- Similarly probability class is difficult goes up from 0.4 to
- $P(d^1|g^3)=0.629$
- If recruiter has lost grade but has letter
- $P(i^1|l^0)=0.14$



	d^0	d^1
	0.6	0.4

	i^0	i^1
	0.7	0.3

	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^0,d^0	0.9	0.08	0.02
i^0,d^1	0.5	0.3	0.2

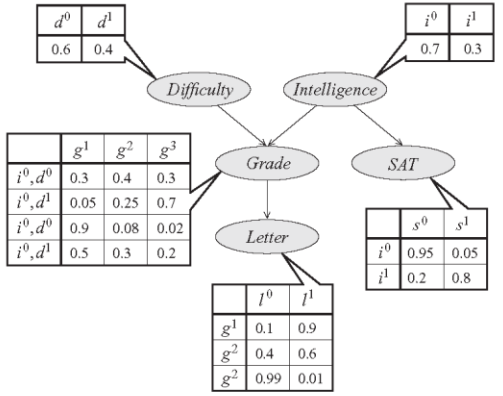
	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^2	0.99	0.01

- Recruiter has both grade and letter
- $P(i^1|l^0,g^3)=0.079$
  - Same as if he had only grade
  - Letter is immaterial
- Reasoning from effects to causes is called evidential reasoning

Intercausal reasoning

- Recruiter has grade (letter does not matter)
- $P(i^1|g^3)=P(i^1|l^0,g^3)=0.079$
- Recruiter receives high SAT score (leads to dramatic increase)
- $P(i^1|g^3,s^1)=0.578$
- Intuition:
  - High SAT score outweighs poor grade since low intelligence rarely gets good SAT scores
  - Smart students more likely to get Cs in hard classes
- Probability of class is difficult also goes up from
- $P(d^1|g^3)=0.629$ to
- $P(d^1|g^3,s^1)=0.76$



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

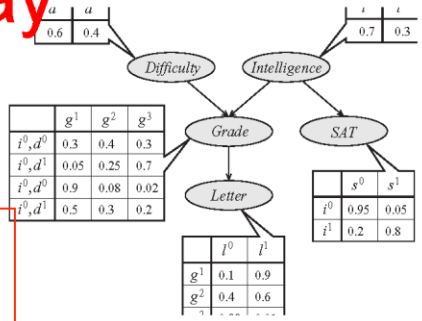
	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^2	0.99	0.01

Information about SAT score gave us information about Intelligence

which with Grade told us about difficulty of course

One causal factor for Grade (Intelligence) gives us information about another (Difficulty)

Explaining Away



An example:

- Given grade
- $P(i^1|l^0,g^3)=0.079$
- If we observe ECON101 is a hard class
- $P(i^1|g^3,d^1)=0.11$
- We have provided partial explanation for George's performance in ECON101

Another example:

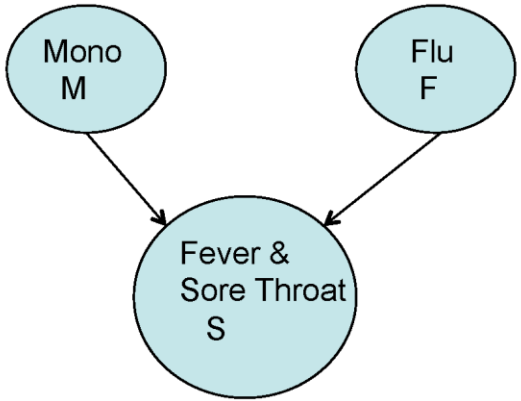
- If George gets a B in ECON101
- $P(i^1|g^2)=0.175$
- If we observe ECON101 is a hard class
- $P(i^1|g^2,d^1)=0.34$
- We have explained away the poor grade via the difficulty of the class

Explaining away is one type of intercausal reasoning

- Different causes of the same effect can interact
- All determined by probability calculation rather than heuristics

Intercausal Reasoning is Common in Human Reasoning

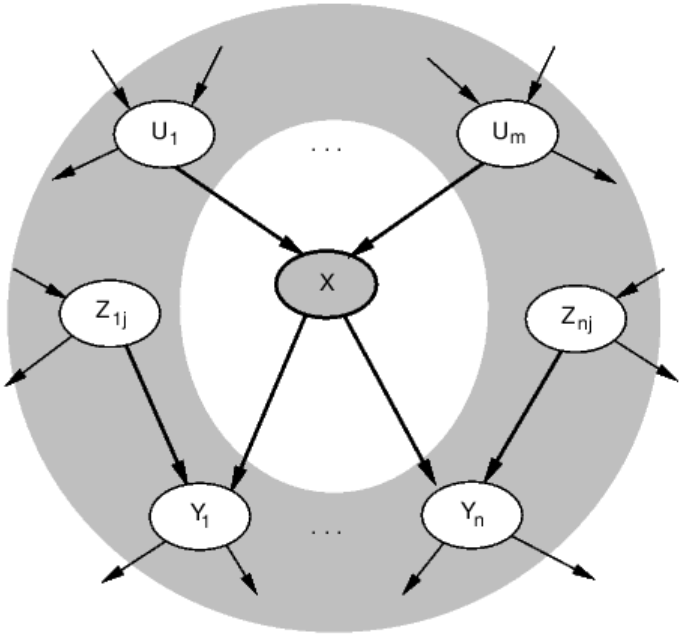
Another example of explaining away



- Binary Variables
- Fever & Sore Throat can be caused by mono and flu
- When flu is diagnosed probability of mono is reduced (although mono could still be present)
- It provides an alternative explanation of symptoms
- $P(m^1|s^1) > P(m^1|s^1,f^1)$

Markov blanket

Each node is conditionally independent of all others given its Markov blanket: parents + children + children's parents



Constructing belief networks

Need a method such that a series of locally testable assertions of conditional independence guarantees the required global semantics

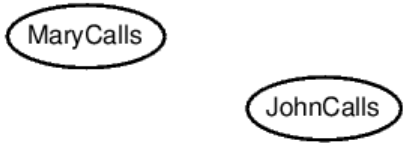
1. Choose an ordering of variables $X_1, \ldots, X_n$
2. For $i = 1$ to n
   add X_i to the network
   select parents from $X_1, \ldots, X_{i-1}$ such that
   $\mathbf{P}(X_i|Parents(X_i)) = \mathbf{P}(X_i|X_1, \ldots, X_{i-1})$

This choice of parents guarantees the global semantics:

$\mathbf{P}(X_1, \ldots, X_n) = \prod_{i=1}^{n} \mathbf{P}(X_i|X_1, \ldots, X_{i-1})$ (chain rule)
$= \prod_{i=1}^{n} \mathbf{P}(X_i|Parents(X_i))$ by construction

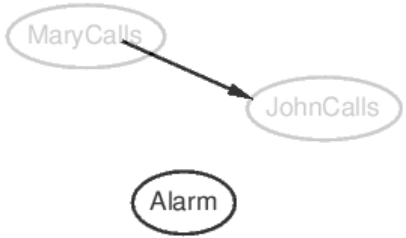
Example

Suppose we choose the ordering M, J, A, B, E



$P(J|M) = P(J)$?

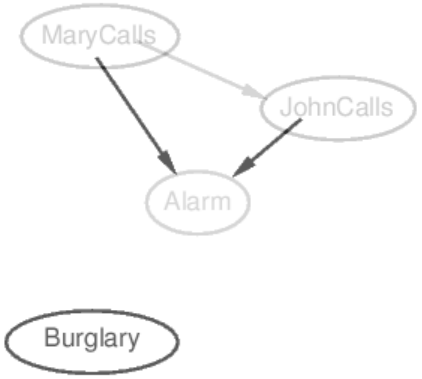
Suppose we choose the ordering M, J, A, B, E



$P(J|M) = P(J)$? No

$P(A|J, M) = P(A|J)$? $P(A|J, M) = P(A)$?

Suppose we choose the ordering M, J, A, B, E



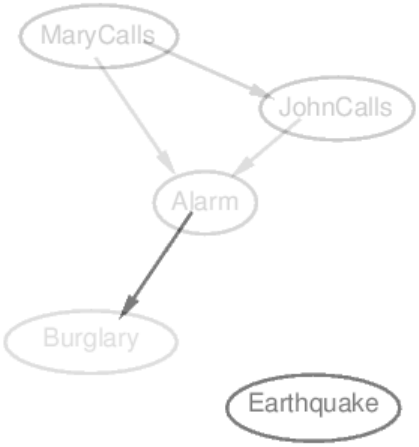
$P(J|M) = P(J)$? No

$P(A|J, M) = P(A|J)$? $P(A|J, M) = P(A)$? No

$P(B|A, J, M) = P(B|A)$?

$P(B|A, J, M) = P(B)$?

Suppose we choose the ordering M, J, A, B, E



$P(J|M) = P(J)$? No

$P(A|J, M) = P(A|J)$? $P(A|J, M) = P(A)$? No

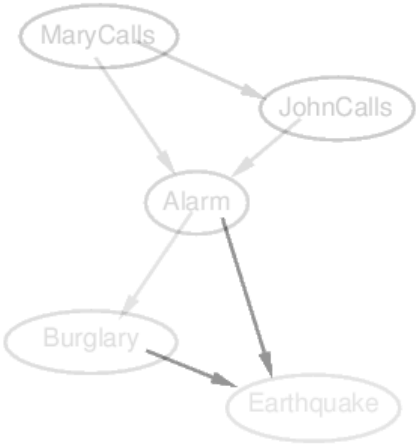
$P(B|A, J, M) = P(B|A)$? Yes

$P(B|A, J, M) = P(B)$? No

$P(E|B, A, J, M) = P(E|A)$?

$P(E|B, A, J, M) = P(E|A, B)$?

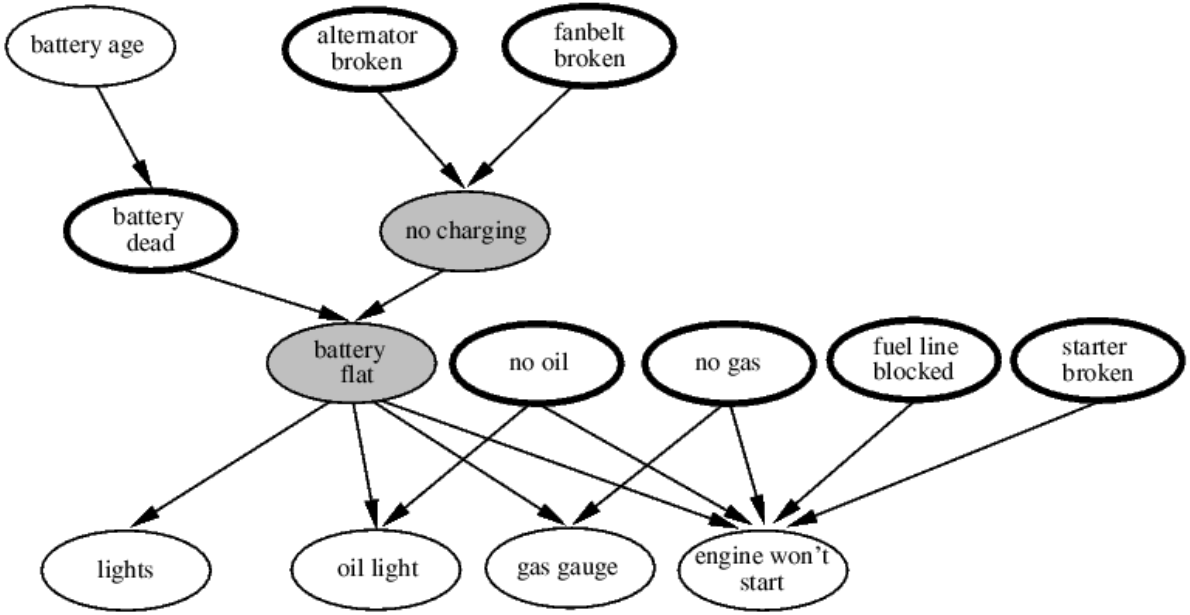
Suppose we choose the ordering M, J, A, B, E



- $P(J|M) = P(J)$? No
- $P(A|J, M) = P(A|J)$? $P(A|J, M) = P(A)$? No
- $P(B|A, J, M) = P(B|A)$? Yes
- $P(B|A, J, M) = P(B)$? No
- $P(E|B, A, J, M) = P(E|A)$? No
- $P(E|B, A, J, M) = P(E|A, B)$? Yes

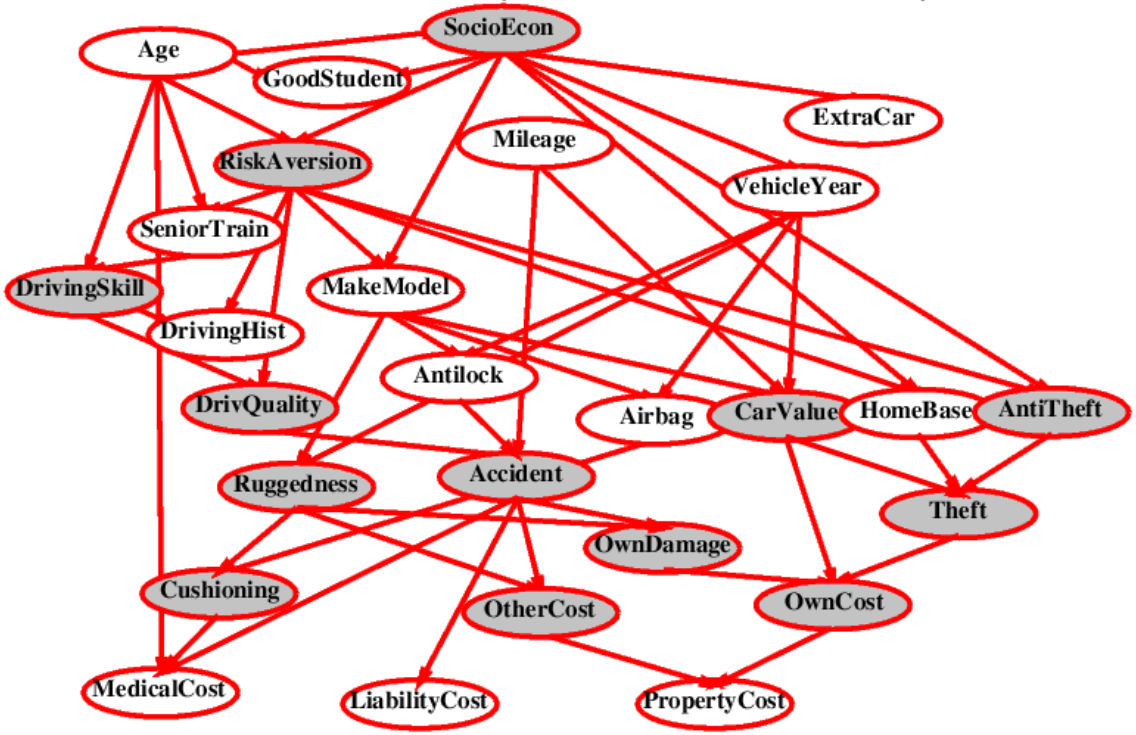
Example: car diagnosis

Initial evidence: engine won't start
Testable variables (thin ovals), diagnosis variables (thick ovals)
Hidden variables (shaded) ensure sparse structure, reduce parameters



Example: car insurance

Predict claim costs (medical, liability, property)
given data on application form (other unshaded nodes)



Compact conditional distributions

CPT grows exponentially with no. of parents
CPT becomes infinite with continuous-valued parent or child

Solution: canonical distributions that are defined compactly

Deterministic nodes are the simplest case:
$X = f(Parents(X))$ for some function f

E.g., Boolean functions
$NorthAmerican \Leftrightarrow Canadian \vee US \vee Mexican$

E.g., numerical relationships among continuous variables

$$\frac{\partial Level}{\partial t} = \text{inflow} + \text{precipation} - \text{outflow} - \text{evaporation}$$

Compact conditional distributions

Noisy-OR distributions model multiple noninteracting causes

1) Parents $U_1 \ldots U_k$ include all causes (can add leak node)
2) Independent failure probability q_i for each cause alone

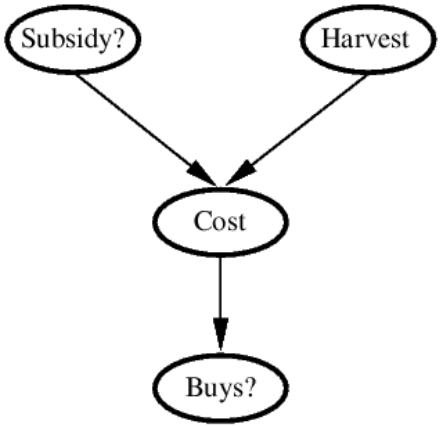
$$\Rightarrow P(X|U_1 \ldots U_j, \neg U_{j+1} \ldots \neg U_k) = 1 - \prod_{i=1}^{j} q_i$$

$Cold$	Flu	$Malaria$	$P(Fever)$	$P(\neg Fever)$
F	F	F	**0.0**	1.0
F	F	T	0.9	**0.1**
F	T	F	0.8	**0.2**
F	T	T	0.98	$0.02 = 0.2 \times 0.1$
T	F	F	0.4	**0.6**
T	F	T	0.94	$0.06 = 0.6 \times 0.1$
T	T	F	0.88	$0.12 = 0.6 \times 0.2$
T	T	T	0.988	$0.012 = 0.6 \times 0.2 \times 0.1$

Number of parameters linear in number of parents

Hybrid (discrete+continuous) networks

Discrete ($Subsidy?$ and $Buys?$); continuous ($Harvest$ and $Cost$)



consumer buys some fruit depending on its cost, which in turn depends on the size of the harvest and whether a government subsidy was received

Option 1: discretization—possibly large errors, large CPTs

Option 2: finitely parameterized canonical families

1) Continuous variable, discrete+continuous parents (e.g., $Cost$)
2) Discrete variable, continuous parents (e.g., $Buys?$)

Continuous child variables

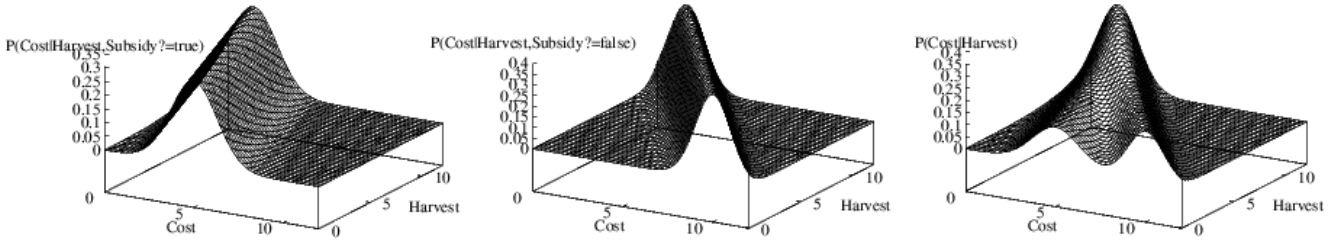
Need one conditional density function for child variable given continuous parents, for each possible assignment to discrete parents

Most common is the linear Gaussian model, e.g.,:

$$
\begin{aligned}
&P(Cost = c | Harvest = h, Subsidy? = true) \\
&= N(a_t h + b_t, \sigma_t)(c) \\
&= \frac{1}{\sigma_t \sqrt{2\pi}} exp\left(-\frac{1}{2}\left(\frac{c - (a_t h + b_t)}{\sigma_t}\right)^2\right)
\end{aligned}
$$

Mean $Cost$ varies linearly with $Harvest$, variance is fixed
Linear variation is unreasonable over the full range
but works OK if the likely range of $Harvest$ is narrow

Continuous child variables



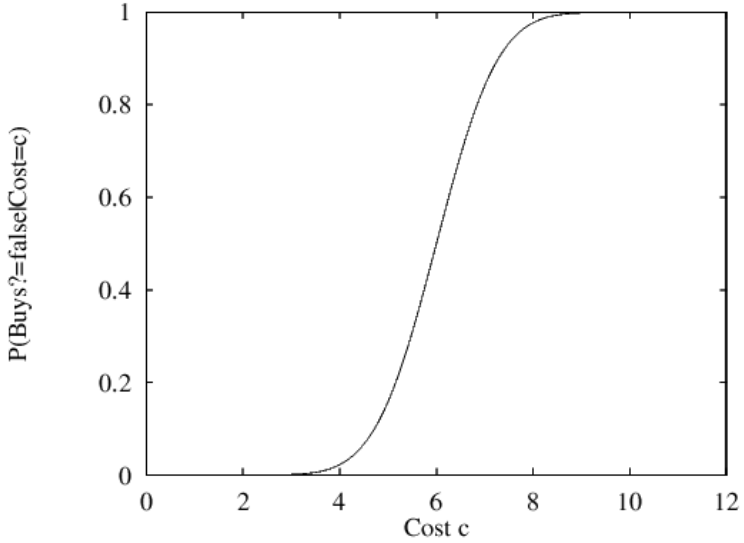
All-continuous network with LG distributions

$\Rightarrow$ full joint is a multivariate Gaussian

Discrete+continuous LG network is a conditional Gaussian network i.e., a multivariate Gaussian over all continuous variables for each combination of discrete variable values

Discrete variable w/ continuous parents

Probability of $Buys?$ given $Cost$ should be a "soft" threshold:



Probit distribution uses integral of Gaussian:

$\Phi(x) = \int_{-\infty}^{x} N(0,1)(x)dx$

$P(Buys? = true \mid Cost = c) = \Phi((-c + \mu)/\sigma)$

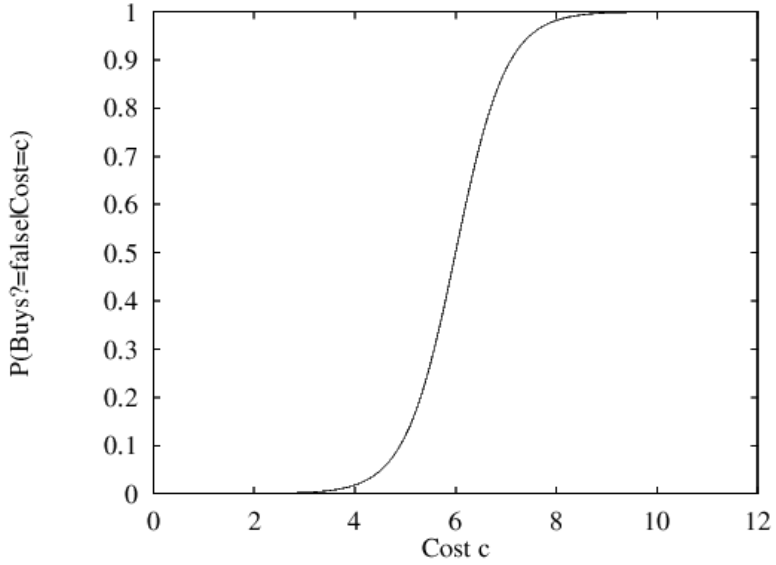
Can view as hard threshold whose location is subject to noise

Discrete variable

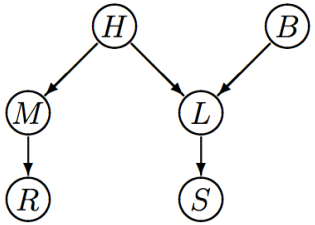
Sigmoid (or logit) distribution also used in neural networks:

$$P(Buys? = true \mid Cost = c) = \frac{1}{1 + exp(-2\frac{-c+\mu}{\sigma})}$$

Sigmoid has similar shape to probit but much longer tails:



Another example



H = hardware problem
B = bug in lab
M = mailer is running
L = lab is running
R = received email
S = problem solved

Hypothesis Evidence

Brute force calculation of $P(H \mid E)$ is done by:

1. Apply the conditional probability rule.

$$P(H \mid E) = P(H \wedge E) \,/\, P(E)$$

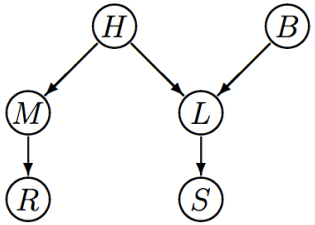
2. Apply the marginal distribution rule to the unknown vertices $\mathbf{U}$.

$$P(H \wedge E) = \sum_{\mathbf{U}=\mathbf{u}} P(H \wedge E \wedge \mathbf{U} = \mathbf{u})$$

3. Apply joint distribution rule for Bayesian networks.

$$\mathbf{P}(X_1, \ldots, X_n) = \prod_{i=1}^{n} \mathbf{P}(X_i \mid Parents(X_i))$$

Another example



H = hardware problem
B = bug in lab
M = mailer is running
L = lab is running
R = received email
S = problem solved

Each node needs a probability table. Size of table depends on number of parents.

$\mathbf{P}(H)$	
$True$	$False$
0.01	0.99

	$\mathbf{P}(M \mid H)$	
H	$True$	$False$
$True$	0.1	0.9
$False$	0.99	0.01

e.g.,

		$\mathbf{P}(L \mid H, B)$	
H	B	$True$	$False$
$True$	$True$	0.01	0.99
$True$	$False$	0.1	0.9
$False$	$True$	0.02	0.98
$False$	$False$	1.0	0.0

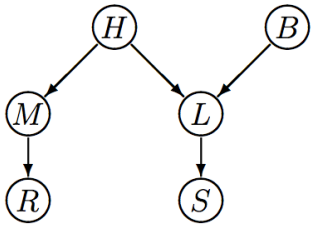
Another example

Calculate $P(B \mid \neg R, S)$ in the buggy lab example.

1. Apply the conditional probability rule.

$$P(B \mid \neg R, S) = \frac{P(B, \neg R, S)}{P(\neg R, S)}$$

2. Apply the marginal distribution rule to the unknown vertices. $P(B, \neg R, S)$ has 3 unknown vertices with $2^3 = 8$ possible value assignments.



H = hardware problem
B = bug in lab
M = mailer is running
L = lab is running
R = received email
S = problem solved

Each node needs a probability table. Size of table depends on number of parents.

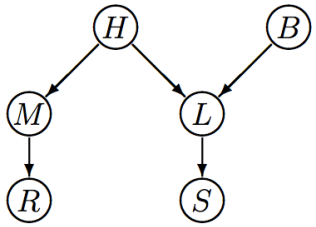
Another example

Calculate $P(B \mid \neg R, S)$ in the buggy lab example.

1. Apply the conditional probability rule.

$$P(B \mid \neg R, S) = \frac{P(B, \neg R, S)}{P(\neg R, S)}$$

2. Apply the marginal distribution rule to the unknown vertices. $P(B, \neg R, S)$ has 3 unknown vertices with $2^3 = 8$ possible value assignments.



H = hardware problem
B = bug in lab
M = mailer is running
L = lab is running
R = received email
S = problem solved

$$\begin{aligned} & P(B, \neg R, S) \\ &= P(B, \neg R, S, H, M, L) \\ &\quad + P(B, \neg R, S, H, M, \neg L) \\ &\quad + P(B, \neg R, S, H, \neg M, L) \\ &\quad + P(B, \neg R, S, H, \neg M, \neg L) \\ &\quad + P(B, \neg R, S, \neg H, M, L) \\ &\quad + P(B, \neg R, S, \neg H, M, \neg L) \\ &\quad + P(B, \neg R, S, \neg H, \neg M, L) \\ &\quad + P(B, \neg R, S, \neg H, \neg M, \neg L) \end{aligned}$$

Another example

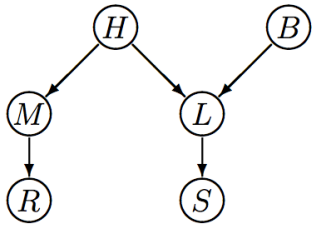
3. Apply joint distribution rule for Bayesian networks. Here are two examples.

$$
\begin{aligned}
&P(B, \neg R, S, H, M, L) \\
&= P(B)\, P(H) \\
&\quad P(M \mid H)\, P(\neg R \mid M) \\
&\quad P(L \mid H, M)\, P(S \mid L)
\end{aligned}
$$

Typo: P(L|H, B)

$$
\begin{aligned}
&P(B, \neg R, S, \neg H, M, \neg L) \\
&= P(B)\, P(\neg H) \\
&\quad P(M \mid \neg H)\, P(\neg R \mid M) \\
&\quad P(\neg L \mid \neg H, M)\, P(S \mid \neg L)
\end{aligned}
$$

Typo: P(-L|-H,B)



H = hardware problem
B = bug in lab
M = mailer is running
L = lab is running
R = received email
S = problem solved