Uncertainty

AIMA2e Chapter 13
Outline

♦ Uncertainty
♦ Probability
♦ Syntax and Semantics
♦ Inference
♦ Independence and Bayes’ Rule
Uncertainty

Let action $A_t = \text{leave for airport } t \text{ minutes before flight}$
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.), Cardano (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* claims of some *probabilistic tendency* in the current situation
(but might be learned from past experience of similar situations)

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:

\[
\begin{align*}
P(A_{25} \text{ gets me there on time}|\ldots) &= 0.04 \\
P(A_{90} \text{ gets me there on time}|\ldots) &= 0.70 \\
P(A_{120} \text{ gets me there on time}|\ldots) &= 0.95 \\
P(A_{1440} \text{ gets me there on time}|\ldots) &= 0.9999
\end{align*}
\]

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
Probability basics

Begin with a set $\Omega$—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 $\Omega$

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

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

\[ P(X = x_i) = \sum_{\{\omega: X(\omega) = x_i\}} P(\omega) \]

e.g., $P(Odd = true) = 1/6 + 1/6 + 1/6 = 1/2$
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 $a = \text{set of sample points where } A(\omega) = \text{true}$
- event $\neg a = \text{set of sample points where } A(\omega) = \text{false}$
- event $a \land b = \text{points where } A(\omega) = \text{true \ and \ } B(\omega) = \text{true}$

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., $A = \text{true}$, $B = \text{false}$, or $a \land \neg b$.

Proposition = disjunction of atomic events in which it is true
- e.g., $(a \lor b) \equiv (\neg a \land b) \lor (a \land \neg b) \lor (a \land b)$

$\Rightarrow P(a \lor b) = P(\neg a \land b) + P(a \land \neg b) + P(a \land b)$
Why use probability?

The definitions imply that certain logically related events must have related probabilities

E.g., \( P(a \lor b) = P(a) + P(b) - P(a \land b) \)

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.
Syntax for propositions

Propositional or Boolean random variables
  e.g., *Cavity* (do I have a cavity?)

Discrete random variables (*finite* or *infinite*)
  e.g., *Weather* is one of \{*sunny*, *rain*, *cloudy*, *snow*\}
  *Weather = rain* is a proposition
  Values must be exhaustive and mutually exclusive

Continuous random variables (*bounded* or *unbounded*)
  e.g., *Temp = 21.6*; also allow, e.g., *Temp < 22.0*.

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

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

Joint probability distribution for a set of r.v.s gives the probability of every atomic event on those r.v.s (i.e., every sample point)
\[
P(Weather, Cavity) = \begin{array}{cccc}
\text{Weather} & \text{sunny} & \text{rain} & \text{cloudy} & \text{snow} \\
\hline
\text{Cavity = true} & 0.144 & 0.02 & 0.016 & 0.02 \\
\text{Cavity = false} & 0.576 & 0.08 & 0.064 & 0.08 \\
\end{array}
\]

Every question about a domain can be answered by the joint distribution because every event is a sum of sample points
Probability for continuous variables

Express distribution as a parameterized function of value:

\[ P(X = x) = U[18, 26](x) = \text{uniform density between 18 and 26} \]

Here \( P \) is a *density*, integrates to 1.

\[ P(X = 20.5) = 0.125 \text{ really means} \]

\[
\lim_{dx \to 0} \frac{P(20.5 \leq X \leq 20.5 + dx)}{dx} = 0.125
\]
Gaussian density

\[ P(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \]
Conditional probability

Conditional or posterior probabilities
  e.g., $P(\text{cavity}|\text{toothache}) = 0.8$
  i.e., given that toothache is all I know
  NOT “if toothache then 80% chance of cavity”

(Notation for conditional distributions:
  $P(Cavity|Toothache) = 2$-element vector of 2-element vectors)

If we know more, e.g., cavity is also given, then we have
  $P(\text{cavity}|\text{toothache}, \text{cavity}) = 1$
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.,
  $P(\text{cavity}|\text{toothache}, \text{49ersWin}) = P(\text{cavity}|\text{toothache}) = 0.8$
This kind of inference, sanctioned by domain knowledge, is crucial
Conditional probability

Definition of conditional probability:

\[ P(a|b) = \frac{P(a \land b)}{P(b)} \text{ if } P(b) \neq 0 \]

**Product rule** gives an alternative formulation:

\[ P(a \land b) = P(a|b)P(b) = p(b|a)P(a) \]

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

\[ P(\text{Weather, Cavity}) = P(\text{Weather}|\text{Cavity})P(\text{Cavity}) \]

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

**Chain rule** is derived by successive application of product rule:

\[
\begin{align*}
P(X_1, \ldots, X_n) &= P(X_1, \ldots, X_{n-1}) P(X_n|X_1, \ldots, X_{n-1}) \\
&= P(X_1, \ldots, X_{n-2}) P(X_{n-1}|X_1, \ldots, X_{n-2}) P(X_n|X_1, \ldots, X_{n-1}) \\
&= \ldots \\
&= \prod_{i=1}^{n} P(X_i|X_1, \ldots, X_{i-1})
\end{align*}
\]
Inference by enumeration

Start with the joint distribution:

<table>
<thead>
<tr>
<th></th>
<th>toothache</th>
<th>¬ toothache</th>
</tr>
</thead>
<tbody>
<tr>
<td>catch</td>
<td>.108</td>
<td>.012</td>
</tr>
<tr>
<td>¬ catch</td>
<td>.072</td>
<td>.008</td>
</tr>
<tr>
<td>cavity</td>
<td>.016</td>
<td>.064</td>
</tr>
<tr>
<td>¬ cavity</td>
<td>.144</td>
<td>.576</td>
</tr>
</tbody>
</table>

For any proposition \( \phi \), 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:

<table>
<thead>
<tr>
<th></th>
<th>toothache</th>
<th>\neg toothache</th>
</tr>
</thead>
<tbody>
<tr>
<td>catch</td>
<td>[.108]</td>
<td>[.012]</td>
</tr>
<tr>
<td>\neg catch</td>
<td>[.072]</td>
<td>[.008]</td>
</tr>
<tr>
<td>cavity</td>
<td>[.016]</td>
<td>[.064]</td>
</tr>
<tr>
<td>\neg cavity</td>
<td>[.144]</td>
<td>[.576]</td>
</tr>
</tbody>
</table>

For any proposition $\phi$, 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:

<table>
<thead>
<tr>
<th></th>
<th>toothache</th>
<th>( \neg \text{toothache} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>catch</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \text{cavity} )</td>
<td>.108</td>
<td>.012</td>
</tr>
<tr>
<td>( \neg \text{cavity} )</td>
<td>.016</td>
<td>.064</td>
</tr>
<tr>
<td>( \text{catch} )</td>
<td>.072</td>
<td>.008</td>
</tr>
<tr>
<td>( \neg \text{catch} )</td>
<td>.144</td>
<td>.576</td>
</tr>
</tbody>
</table>

For any proposition \( \phi \), sum the atomic events where it is true:

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

\[
P(\text{cavity} \lor \text{toothache}) = 0.108 + 0.012 + 0.072 + 0.008 + 0.016 + 0.064 = 0.28
\]
Inference by enumeration

Start with the joint distribution:

<table>
<thead>
<tr>
<th></th>
<th>toothache</th>
<th>¬ toothache</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>catch</td>
<td>¬ catch</td>
</tr>
<tr>
<td>cavity</td>
<td>.108</td>
<td>.012</td>
</tr>
<tr>
<td>¬ cavity</td>
<td>.016</td>
<td>.064</td>
</tr>
</tbody>
</table>

Can also compute conditional probabilities:

\[
P(\neg \text{cavity} | \text{toothache}) = \frac{P(\neg \text{cavity} \land \text{toothache})}{P(\text{toothache})} = \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064} = 0.4
\]
## Normalization

<table>
<thead>
<tr>
<th></th>
<th>toothache</th>
<th>¬ toothache</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>catch</td>
<td>¬ catch</td>
</tr>
<tr>
<td>cavity</td>
<td>.108</td>
<td>.012</td>
</tr>
<tr>
<td>¬ cavity</td>
<td>.016</td>
<td>.064</td>
</tr>
</tbody>
</table>

Denominator can be viewed as a *normalization constant* $\alpha$

\[
P(C\text{avity}|\text{toothache}) = \alpha P(C\text{avity},\text{toothache})
= \alpha [P(C\text{avity},\text{toothache},\text{catch}) + P(C\text{avity},\text{toothache},\neg \text{catch})]
= \alpha [\langle 0.108, 0.016 \rangle + \langle 0.012, 0.064 \rangle]
= \alpha \langle 0.12, 0.08 \rangle = \langle 0.6, 0.4 \rangle
\]

General idea: compute distribution on query variable by fixing *evidence variables* and summing over *hidden variables*
Inference by enumeration, contd.

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

Let the hidden variables be $H = X - Y - E$

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

$$P(Y|E=e) = \alpha P(Y, E=e) = \alpha \sum_h P(Y, E=e, H=h)$$

The terms in the summation are joint entries because $Y$, $E$, and $H$ 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???
Independence

\[ A \text{ and } B \text{ are independent iff} \]
\[ P(A|B) = P(A) \quad \text{or} \quad P(B|A) = P(B) \quad \text{or} \quad P(A, B) = P(A)P(B) \]

\[ P(\text{Toothache, Catch, Cavity, Weather}) = P(\text{Toothache, Catch, Cavity})P(\text{Weather}) \]

32 entries reduced to 12; for \( n \) independent biased coins, \( 2^n \rightarrow n \)

Absolute independence powerful but rare

Dentistry is a large field with hundreds of variables, none of which are independent. What to do?
**Conditional independence**

\[ P(Toothache, Cavity, Catch) \] 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:

\[ (1) \quad P(\text{catch}|\text{toothache}, \text{cavity}) = P(\text{catch}|\text{cavity}) \]

The same independence holds if I haven’t got a cavity:

\[ (2) \quad P(\text{catch}|\text{toothache}, \neg \text{cavity}) = P(\text{catch}|\neg \text{cavity}) \]

* Catch is *conditionally independent* of Toothache given Cavity:  
  \[ P(\text{Catch}|\text{Toothache}, \text{Cavity}) = P(\text{Catch}|\text{Cavity}) \]

Equivalent statements:

\[ P(\text{Toothache}|\text{Catch}, \text{Cavity}) = P(\text{Toothache}|\text{Cavity}) \]
\[ P(\text{Toothache}, \text{Catch}|\text{Cavity}) = P(\text{Toothache}|\text{Cavity})P(\text{Catch}|\text{Cavity}) \]
Conditional independence contd.

Write out full joint distribution using chain rule:
\[
P(Toothache, Catch, Cavity) \\
= P(Toothache | Catch, Cavity) P(Catch, Cavity) \\
= P(Toothache | Catch, Cavity) P(Catch | Cavity) P(Cavity) \\
= P(Toothache | Cavity) P(Catch | Cavity) P(Cavity)
\]

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

In most cases, the use of conditional independence reduces the size of the representation of the joint distribution from exponential in \(n\) to linear in \(n\).

Conditional independence is our most basic and robust form of knowledge about uncertain environments.
Bayes’ Rule

Product rule \( P(a \land b) = P(a|b)P(b) = P(b|a)P(a) \)

\[ \Rightarrow \text{Bayes’ rule } P(a|b) = \frac{P(b|a)P(a)}{P(b)} \]

or in distribution form

\[ P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} = \alpha P(X|Y)P(Y) \]

Useful for assessing diagnostic probability from causal probability:

\[ P(\text{Cause}|\text{Effect}) = \frac{P(\text{Effect}|\text{Cause})P(\text{Cause})}{P(\text{Effect})} \]

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!
Bayes’ Rule and conditional independence

\[
\begin{align*}
P(Cavity | \text{toothache} \land \text{catch}) \\
&= \alpha P(\text{toothache} \land \text{catch} | Cavity) P(Cavity) \\
&= \alpha P(\text{toothache} | Cavity) P(\text{catch} | Cavity) P(Cavity)
\end{align*}
\]

This is an example of a \textit{naive Bayes} model:

\[
P(\text{Cause}, \text{Effect}_1, \ldots, \text{Effect}_n) = P(\text{Cause}) \Pi_i P(\text{Effect}_i | \text{Cause})
\]

Total number of parameters is \textit{linear} in \( n \)
Summary

Probability is a rigorous formalism for uncertain knowledge

Joint probability distribution specifies probability of every atomic event

Queries can be answered by summing over atomic events

For nontrivial domains, we must find a way to reduce the joint size

Independence and conditional independence provide the tools
BAYESIAN NETWORKS

AIMA2e CHAPTER 14.1–3
Outline

◊ Syntax
◊ Semantics
◊ Parameterized distributions
Bayesian 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:
  \[ P(X_i|Parents(X_i)) \]

In the simplest case, conditional distribution represented as a conditional probability table (CPT) giving the distribution over \( X_i \) for each combination of parent values
Example

Topology of network encodes conditional independence assertions:

- **Weather** is independent of the other variables
- **Toothache** and **Catch** are conditionally independent given **Cavity**
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:

- A burglar can set the alarm off
- An earthquake can set the alarm off
- The alarm can cause Mary to call
- The alarm can cause John to call
Example contd.

| B | E | P(A|B,E) |
|---|---|---------|
| T | T | .95     |
| T | F | .94     |
| F | T | .29     |
| F | F | .001    |

| A | P(J|A) |
|---|------|
| T | .90  |
| F | .05  |

| A | P(M|A) |
|---|------|
| T | .70  |
| F | .01  |
Compactness

A CPT for Boolean $X_i$ with $k$ Boolean parents has $2^k$ rows for the combinations of parent values.

Each row requires one number $p$ for $X_i = true$ (the number for $X_i = false$ is just $1 - p$).

If each variable has no more than $k$ parents, the complete network requires $O(n \cdot 2^k)$ numbers.

I.e., grows linearly with $n$, vs. $O(2^n)$ for the full joint distribution.

For burglary net, $1 + 1 + 4 + 2 + 2 = 10$ numbers (vs. $2^5 - 1 = 31$)
Global semantics defines the full joint distribution as the product of the local conditional distributions:

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

e.g., $P(j \land m \land a \land \neg b \land \neg e)$

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

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

e.g., $P(j \land m \land a \land \neg b \land \neg e)$

$$= P(j|a)P(m|a)P(a|\neg b, \neg e)P(\neg b)P(\neg e)$$
Local semantics: each node is conditionally independent of its nondescendants given its parents

Theorem: Local semantics $\iff$ global semantics
Markov blanket

Each node is conditionally independent of all others given its
Markov blanket: parents + children + children’s parents
Constructing Bayesian 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
     $$ P(X_i|Parents(X_i)) = P(X_i|X_1, \ldots, X_{i-1}) $$

This choice of parents guarantees the global semantics:

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

$$ = \prod_{i=1}^{n} P(X_i|Parents(X_i)) $$ (by construction)
Example

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

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

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

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

Suppose we choose the ordering \( M, J, A, B, E \)

\[
\begin{align*}
P(J|M) &= P(J)? \quad \text{No} \\
P(A|J, M) &= P(A|J)? \quad P(A|J, M) = P(A)? \quad \text{No} \\
P(B|A, J, M) &= P(B|A)? \\
P(B|A, J, M) &= P(B)?
\end{align*}
\]
Example

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

![Diagram showing the relationships between MaryCalls, JohnCalls, Alarm, Burglary, and Earthquake]

- $P(J|M) = P(J) \quad \text{No}$
- $P(A|J, M) = P(A|J) \quad P(A|J, M) = P(A) \quad \text{No}$
- $P(B|A, J, M) = P(B|A) \quad \text{Yes}$
- $P(B|A, J, M) = P(B) \quad \text{No}$
- $P(E|B, A, J, M) = P(E|A) \quad P(E|B, A, J, M) = P(E|A, B)$
Example

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

$P(J|M) = P(J)$?  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
Deciding conditional independence is hard in noncausal directions
(Causal models and conditional independence seem hardwired for humans!)
Assessing conditional probabilities is hard in noncausal directions
Network is less compact: $1 + 2 + 4 + 2 + 4 = 13$ numbers needed
Example: Car diagnosis

Initial evidence: car won’t start
Testable variables (green), “broken, so fix it” variables (orange)
Hidden variables (gray) ensure sparse structure, reduce parameters
Example: Car insurance
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:
\[ X = f(\text{Parents}(X)) \] for some function \( f \)

E.g., Boolean functions
\[ \text{NorthAmerican} \iff \text{Canadian} \lor \text{US} \lor \text{Mexican} \]

E.g., numerical relationships among continuous variables
\[ \frac{\partial \text{Level}}{\partial t} = \text{inflow} + \text{precipitation} - \text{outflow} - \text{evaporation} \]
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

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

<table>
<thead>
<tr>
<th>Cold</th>
<th>Flu</th>
<th>Malaria</th>
<th>$P(\text{Fever})$</th>
<th>$P(\neg \text{Fever})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>F</td>
<td>F</td>
<td>0.0</td>
<td>1.0</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
<td>0.9</td>
<td>0.1</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>F</td>
<td>0.8</td>
<td>0.2</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>T</td>
<td>0.98</td>
<td>0.02 = 0.2 $\times$ 0.1</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>F</td>
<td>0.4</td>
<td>0.6</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>T</td>
<td>0.94</td>
<td>0.06 = 0.6 $\times$ 0.1</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>F</td>
<td>0.88</td>
<td>0.12 = 0.6 $\times$ 0.2</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>0.988</td>
<td>0.012 = 0.6 $\times$ 0.2 $\times$ 0.1</td>
</tr>
</tbody>
</table>

Number of parameters **linear** in number of parents
Hybrid (discrete+continuous) networks

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

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.:

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

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

⇒ full joint distribution 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 $\textit{Buys?}$ given $\textit{Cost}$ should be a “soft” threshold:

![Graph showing the relationship between cost and the probability of Buys?](image)

**Probit** distribution uses integral of Gaussian:

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

$$
P(\text{Buys?=true} \mid \text{Cost}=c) = \Phi((-c + \mu)/\sigma)
$$
Why the probit?

1. It’s sort of the right shape

2. Can view as hard threshold whose location is subject to noise

![Diagram showing the relationship between Cost, Noise, and Buys?]
Discrete variable contd.

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

\[ P(\text{Buys}^? = \text{true} \mid \text{Cost} = c) = \frac{1}{1 + e^{\frac{-c+\mu}{\sigma}}} \]

Sigmoid has similar shape to probit but much longer tails:
Summary

Bayes nets provide a natural representation for (causally induced) conditional independence

Topology + CPTs = compact representation of joint distribution

Generally easy for (non)experts to construct

Canonical distributions (e.g., noisy-OR) = compact representation of CPTs

Continuous variables ⇒ parameterized distributions (e.g., linear Gaussian)
Inference in Bayesian networks

AIMA2e Chapter 14.4–5
Outline

◊ Exact inference by enumeration
◊ Exact inference by variable elimination
◊ Approximate inference by stochastic simulation
◊ Approximate inference by Markov chain Monte Carlo
**Inference tasks**

Simple queries: compute posterior marginal $P(X_i|E=e)$
   e.g., $P(\text{NoGas}|\text{Gauge = empty, Lights = on, Starts = false})$

Conjunctive queries: $P(X_i, X_j|E=e) = P(X_i|E=e)P(X_j|X_i, E=e)$

Optimal decisions: decision networks include utility information;
   probabilistic inference required for $P(\text{outcome}|\text{action, evidence})$

Value of information: which evidence to seek next?

Sensitivity analysis: which probability values are most critical?

Explanation: why do I need a new starter motor?
Inference by enumeration

Slightly intelligent way to sum out variables from the joint without actually constructing its explicit representation

Simple query on the burglary network:
\[ P(B|j, m) \]
\[ = P(B, j, m)/P(j, m) \]
\[ = \alpha P(B, j, m) \]
\[ = \alpha \sum_e \sum_a P(B, e, a, j, m) \]

Rewrite full joint entries using product of CPT entries:
\[ P(B|j, m) \]
\[ = \alpha \sum_e \sum_a P(B) P(e) P(a|B, e) P(j|a) P(m|a) \]
\[ = \alpha P(B) \sum_e P(e) \sum_a P(a|B, e) P(j|a) P(m|a) \]

Recursive depth-first enumeration: \( O(n) \) space, \( O(d^n) \) time
**Enumeration algorithm**

function `ENUMERATION-ASK(X, e, bn)` returns a distribution over $X$

**inputs:**
- $X$, the query variable
- $e$, observed values for variables $E$
- $bn$, a Bayesian network with variables $\{X\} \cup E \cup Y$

$Q(X) \leftarrow$ a distribution over $X$, initially empty
for each value $x_i$ of $X$ do
    extend $e$ with value $x_i$ for $X$
    $Q(x_i) \leftarrow $ `ENUMERATE-ALL(VARS[bn], e)``
return `NORMALIZE(Q(X))`

function `ENUMERATE-ALL(vars, e)` returns a real number

if `EMPTY?(vars)` then return 1.0

$Y \leftarrow $ `FIRST(vars)`
if $Y$ has value $y$ in $e$
    then return $P(y \mid Pa(Y)) \times $ `ENUMERATE-ALL(REST(vars), e)`
else return $\sum_y P(y \mid Pa(Y)) \times $ `ENUMERATE-ALL(REST(vars), e_y)`
    where $e_y$ is $e$ extended with $Y = y$
Enumeration is inefficient: repeated computation
e.g., computes $P(j|a)P(m|a)$ for each value of $e$
Inference by variable elimination

Variable elimination: carry out summations right-to-left, storing intermediate results (factors) to avoid recomputation

\[
P(B|j, m) = \alpha \frac{P(B) \sum_e P(e) \sum_a P(a|B, e) P(j|a) P(m|a)}{P(B) \sum_e P(e) \sum_a P(a|B, e) P(j|a) f_M(a)}
\]

\[
= \alpha P(B) \sum_e P(e) \sum_a P(a|B, e) f_J(a) f_M(a)
\]

\[
= \alpha P(B) \sum_e P(e) \sum_a f_A(a, b, e) f_J(a) f_M(a)
\]

\[
= \alpha P(B) \sum_e P(e) f_{AJM}(b, e) \text{ (sum out } A) \]

\[
= \alpha P(B) f_{EAJM}(b) \text{ (sum out } E) \]

\[
= \alpha f_B(b) \times f_{EAJM}(b)
\]
Variable elimination: Basic operations

Summing out a variable from a product of factors:
move any constant factors outside the summation
add up submatrices in pointwise product of remaining factors

\[ \sum_x f_1 \times \cdots \times f_k = f_1 \times \cdots \times f_i \sum_x f_{i+1} \times \cdots \times f_k = f_1 \times \cdots \times f_i \times f_X \]

assuming \( f_1, \ldots, f_i \) do not depend on \( X \)

Pointwise product of factors \( f_1 \) and \( f_2 \):
\[
f_1(x_1, \ldots, x_j, y_1, \ldots, y_k) \times f_2(y_1, \ldots, y_k, z_1, \ldots, z_l) = f(x_1, \ldots, x_j, y_1, \ldots, y_k, z_1, \ldots, z_l)
\]

E.g., \( f_1(a, b) \times f_2(b, c) = f(a, b, c) \)
**Variable elimination algorithm**

function ELIMINATION-ASK(X, e, bn) returns a distribution over X

inputs: X, the query variable
        e, evidence specified as an event
        bn, a belief network specifying joint distribution \( P(X_1, \ldots, X_n) \)

factors ← []; vars ← REVERSE(VARS[bn])

for each var in vars do
    factors ← [MAKE-FACTOR(var, e)]|factors|
    if var is a hidden variable then factors ← SUM-OUT(var, factors)

return NORMALIZE(POINTWISE-PRODUCT(factors))
Irrelevant variables

Consider the query \( P(JohnCalls|Burglary = true) \)

\[
P(J|b) = \alpha P(b) \sum_e P(e) \sum_a P(a|b, e) P(J|a) \sum_m P(m|a)
\]

Sum over \( m \) is identically 1; \( M \) is irrelevant to the query

Thm 1: \( Y \) is irrelevant unless \( Y \in Ancestors(\{X\} \cup E) \)

Here, \( X = JohnCalls, E = \{Burglary\} \), and
\( Ancestors(\{X\} \cup E) = \{Alarm, Earthquake\} \)
so \( M \) is irrelevant

(Compare this to backward chaining from the query in Horn clause KBs)
Irrelevant variables contd.

Defn: **moral graph** of Bayes net: marry all parents and drop arrows

Defn: **A** is m-separated from **B** by **C** iff separated by **C** in the moral graph

Thm 2: **Y** is irrelevant if m-separated from **X** by **E**

For $P(\text{JohnCalls} | \text{Alarm} = \text{true})$, both

**Burglary** and **Earthquake** are irrelevant
Complexity of exact inference

Singly connected networks (or polytrees):
- any two nodes are connected by at most one (undirected) path
- time and space cost of variable elimination are $O(d^k n)$

Multiply connected networks:
- can reduce 3SAT to exact inference \( \Rightarrow \) NP-hard
- equivalent to \textit{counting} 3SAT models \( \Rightarrow \) \#P-complete

1. A $\lor$ B $\lor$ C
2. C $\lor$ D $\lor$ $\neg$A
3. B $\lor$ C $\lor$ $\neg$D
Inference by stochastic simulation

Basic idea:
1) Draw $N$ samples from a sampling distribution $S$
2) Compute an approximate posterior probability $\hat{P}$
3) Show this converges to the true probability $P$

Outline:
- Sampling from an empty network
- Rejection sampling: reject samples disagreeing with evidence
- Likelihood weighting: use evidence to weight samples
- Markov chain Monte Carlo (MCMC): sample from a stochastic process whose stationary distribution is the true posterior
Sampling from an empty network

function PRIOR-SAMPLE($bn$) returns an event sampled from $bn$
inputs: $bn$, a belief network specifying joint distribution $P(X_1, \ldots, X_n)$

$x$ ← an event with $n$ elements
for $i = 1$ to $n$ do
    $x_i$ ← a random sample from $P(X_i \mid Parents(X_i))$
return $x$
Example

Cloudy

| C | P(S|C) |
|---|-------|
| T | .10   |
| F | .50   |

Sprinkler

| S | R | P(W|S,R) |
|---|---|---------|
| T | T | .99     |
| T | F | .90     |
| F | T | .90     |
| F | F | .01     |

Wet Grass

Rain

| C | P(R|C) |
|---|-------|
| T | .80   |
| F | .20   |

P(C)

.50

=>?

AB C D E F A G T I J I K L T L
Example

| C | P(S|C) |
|---|-------|
| T | .10   |
| F | .50   |

| S | R | P(W|S,R) |
|---|--|---------|
| T | T | .99     |
| T | F | .90     |
| F | T | .90     |
| F | F | .01     |

| C | P(R|C) |
|---|-------|
| T | .80   |
| F | .20   |

Cloudy

Sprinkler

Wet Grass

Rain

\[ P(C) = 0.50 \]

\[ P(S|C) = \begin{array}{cc} 0.10 & 0.50 \\ \end{array} \]

\[ P(W|S,R) = \begin{array}{cccc} 0.99 & 0.90 & 0.90 & 0.01 \\ \\ \end{array} \]

\[ P(R|C) = \begin{array}{c} 0.80 \\ 0.20 \\ \end{array} \]
Example

- **P(C)**
  - 0.50

| C | P(S|C) |
|---|---|
| T | 0.10 |
| F | 0.50 |

| C | P(R|C) |
|---|---|
| T | 0.80 |
| F | 0.20 |

| S | R | P(W|S,R) |
|---|---|---|
| T | T | 0.99 |
| T | F | 0.90 |
| F | T | 0.90 |
| F | F | 0.01 |
Example

Cloudy

| C | P(S|C) |
|---|-------|
| T | .10   |
| F | .50   |

Sprinkler

| S | R | P(W|S,R) |
|---|---|---------|
| T | T | .99     |
| T | F | .90     |
| F | T | .90     |
| F | F | .01     |

Rain

| C | P(R|C) |
|---|-------|
| T | .80   |
| F | .20   |

Wet Grass

P(C) = .50
Cloudy

Sprinkler

Wet Grass

Rain

| C | P(S|C) |
|---|-------|
| T | .10   |
| F | .50   |

| C | P(R|C) |
|---|-------|
| T | .80   |
| F | .20   |

| S | R | P(W|S,R) |
|---|---|---------|
| T | T | .99     |
| T | F | .90     |
| F | T | .90     |
| F | F | .01     |

Example
### Example

The diagram represents a Bayesian network with the following variables and probabilities:

- **Cloudy**
  - P(C) = 0.50
  - P(S|R) = 0.90
  - P(R|C) = 0.80

- **Sprinkler**
  - P(S|C) =
    - T: 0.10
    - F: 0.50

- **Rain**
  - P(R|C) =
    - T: 0.80
    - F: 0.20

- **Wet Grass**
  - P(W|S,R) =
    - T T: 0.99
    - T F: 0.90
    - F T: 0.90
    - F F: 0.01
Example

- **Cloudy**
  - P(C) = 0.50

- **Sprinkler**
  - P(S|C):
    - T: 0.10
    - F: 0.50

- **Rain**
  - P(R|C):
    - T: 0.80
    - F: 0.20

- **Wet Grass**
  - P(W|S,R):
    - T T: 0.99
    - T F: 0.90
    - F T: 0.90
    - F F: 0.01
Sampling from an empty network contd.

Probability that PriorSample generates a particular event

\[ S_{PS}(x_1 \ldots x_n) = \prod_{i=1}^{n} P(x_i|Parents(X_i)) = P(x_1 \ldots x_n) \]

i.e., the true prior probability

E.g., \[ S_{PS}(t, f, t, t) = 0.5 \times 0.9 \times 0.8 \times 0.9 = 0.324 = P(t, f, t, t) \]

Let \( N_{PS}(x_1 \ldots x_n) \) be the number of samples generated for event \( x_1, \ldots, x_n \)

Then we have

\[
\lim_{N \to \infty} \hat{P}(x_1, \ldots, x_n) = \lim_{N \to \infty} \frac{N_{PS}(x_1, \ldots, x_n)}{N} \\
= S_{PS}(x_1, \ldots, x_n) \\
= P(x_1 \ldots x_n)
\]

That is, estimates derived from PriorSample are consistent

Shorthand: \( \hat{P}(x_1, \ldots, x_n) \approx P(x_1 \ldots x_n) \)
Rejection sampling

\( \hat{P}(X|e) \) estimated from samples agreeing with \( e \)

```
function REJECTION-SAMPLING(X, e, bn, N) returns an estimate of \( P(X|e) \)
    local variables: N, a vector of counts over X, initially zero
    for j = 1 to N do
        x ← PRIOR-SAMPLE(bn)
        if x is consistent with e then
            N[x] ← N[x] + 1 where \( x \) is the value of \( X \) in \( x \)
    return normalize(N[X])
```

E.g., estimate \( P(Rain|Sprinkler = true) \) using 100 samples

27 samples have \( Sprinkler = true \)

Of these, 8 have \( Rain = true \) and 19 have \( Rain = false \).

\[ \hat{P}(Rain|Sprinkler = true) = \text{normalize}(\langle 8, 19 \rangle) = \langle 0.296, 0.704 \rangle \]

Similar to a basic real-world empirical estimation procedure
Analysis of rejection sampling

\[ \hat{P}(X|e) = \alpha N_{PS}(X, e) \quad \text{(algorithm defn.)} \]
\[ = \frac{N_{PS}(X, e)}{N_{PS}(e)} \quad \text{(normalized by } N_{PS}(e)) \]
\[ \approx \frac{P(X, e)}{P(e)} \quad \text{(property of PriorsSample)} \]
\[ = P(X|e) \quad \text{(defn. of conditional probability)} \]

Hence rejection sampling returns consistent posterior estimates

Problem: hopelessly expensive if \( P(e) \) is small

\( P(e) \) drops off exponentially with number of evidence variables!
Likelihood weighting

Idea: fix evidence variables, sample only nonevidence variables, and weight each sample by the likelihood it accords the evidence.

```plaintext
function LIKELIHOOD-WEIGHTING(X, e, bn, N) returns an estimate of P(X|e)
  local variables: W, a vector of weighted counts over X, initially zero
    for j = 1 to N do
      x, w ← WEIGHTED-SAMPLE(bn)
      W[x] ← W[x] + w where x is the value of X in x
    return NORMALIZE(W[X])

function WEIGHTED-SAMPLE(bn, e) returns an event and a weight
  x ← an event with n elements; w ← 1
  for i = 1 to n do
    if X_i has a value x_i in e
      then w ← w × P(X_i = x_i | Parents(X_i))
        else x_i ← a random sample from P(X_i | Parents(X_i))
  return x, w
```

\[ w = 1.0 \]
\begin{itemize}
\item P(C) = 0.50
\item P(S|C) = \begin{tabular}{|c|c|}
\hline
C & P(S|C) \\
\hline
T & 0.10 \\
F & 0.50 \\
\hline
\end{tabular}
\item P(W|S,R) = \begin{tabular}{|c|c|c|}
\hline
S & R & P(W|S,R) \\
\hline
T & T & 0.99 \\
T & F & 0.90 \\
F & T & 0.90 \\
F & F & 0.01 \\
\hline
\end{tabular}
\item P(R|C) = \begin{tabular}{|c|c|}
\hline
C & P(R|C) \\
\hline
T & 0.80 \\
F & 0.20 \\
\hline
\end{tabular}
\end{itemize}

\textit{w} = 1.0
Likelihood weighting example

\[
P(C) = 0.50
\]

\[
\begin{array}{c|c|c|c|c|}
C & P(S|C) & P(W|S,R) & & \\
T & 0.10 & 0.99 & & \\
F & 0.50 & 0.90 & & \\
\end{array}
\]

\[
\begin{array}{c|c}
C & P(R|C) \\
T & 0.80 \\
F & 0.20 \\
\end{array}
\]

\[
w = 1.0
\]
\[ w = 1.0 \times 0.1 \]
\[ w = 1.0 \times 0.1 \]
Likelihood weighting example

\[
\begin{array}{c|c}
C & P(S|C) \\
\hline
T & .10 \\
F & .50 \\
\end{array}
\]

\[
\begin{array}{c|c}
C & P(R|C) \\
\hline
T & .80 \\
F & .20 \\
\end{array}
\]

\[
\begin{array}{c|c|c}
S & R & P(W|S,R) \\
\hline
T & T & .99 \\
T & F & .90 \\
F & T & .90 \\
F & F & .01 \\
\end{array}
\]

\[w = 1.0 \times 0.1\]
\[ w = 1.0 \times 0.1 \times 0.99 = 0.099 \]
Likelihood weighting analysis

Sampling probability for \texttt{WEIGHTEDSAMPLE} is

$$S_{WS}(z, e) = \prod_{i=1}^{l} P(z_i \mid \text{Parents}(Z_i))$$

Note: pays attention to evidence in \texttt{ancestors} only

$$\Rightarrow$$ somewhere “in between” prior and posterior distribution

Weight for a given sample \(z, e\) is

$$w(z, e) = \prod_{i=1}^{m} P(e_i \mid \text{Parents}(E_i))$$

Weighted sampling probability is

$$S_{WS}(z, e)w(z, e)$$

$$= \prod_{i=1}^{l} P(z_i \mid \text{Parents}(Z_i)) \cdot \prod_{i=1}^{m} P(e_i \mid \text{Parents}(E_i))$$

$$= P(z, e) \quad \text{(by standard global semantics of network)}$$

Hence likelihood weighting returns consistent estimates
but performance still degrades with many evidence variables
because a few samples have nearly all the total weight
Approximate inference using MCMC

“State” of network = current assignment to all variables.

Generate next state by sampling one variable given Markov blanket
Sample each variable in turn, keeping evidence fixed

```
function MCMC-Ask(X, e, bn, N) returns an estimate of P(X|e)
    local variables: N[X], a vector of counts over X, initially zero
    Z, the nonevidence variables in bn
    x, the current state of the network, initially copied from e

    initialize x with random values for the variables in Y
    for j = 1 to N do
        N[x] ← N[x] + 1 where x is the value of X in x
        for each Z_i in Z do
            sample the value of Z_i in x from P(Z_i|MB(Z_i)) given the values of
            MB(Z_i) in x
        return Normalize(N[X])
```

Can also choose a variable to sample at random each time
The Markov chain

With $Sprinkler = true, WetGrass = true$, there are four states:

Wander about for a while, average what you see
MCMC example contd.

Estimate $\mathbf{P}(\text{Rain}|\text{Sprinkler} = \text{true}, \text{WetGrass} = \text{true})$

Sample $\text{Cloudy}$ or $\text{Rain}$ given its Markov blanket, repeat. Count number of times $\text{Rain}$ is true and false in the samples.

E.g., visit 100 states
- 31 have $\text{Rain} = \text{true}$, 69 have $\text{Rain} = \text{false}$

$\hat{\mathbf{P}}(\text{Rain}|\text{Sprinkler} = \text{true}, \text{WetGrass} = \text{true})$

$= \text{NORMALIZE}(\langle 31, 69 \rangle) = \langle 0.31, 0.69 \rangle$

Theorem: chain approaches stationary distribution:
- long-run fraction of time spent in each state is exactly proportional to its posterior probability
Markov blanket sampling

Markov blanket of *Cloudy* is
*Sprinkler* and *Rain*

Markov blanket of *Rain* is
*Cloudy, Sprinkler, and WetGrass*

Probability given the Markov blanket is calculated as follows:

\[ P(x'_i|MB(X_i)) = P(x'_i|Parents(X_i)) \prod_{Z_j \in \text{Children}(X_i)} P(z_j|Parents(Z_j)) \]

Easily implemented in message-passing parallel systems, brains

Main computational problems:
1) Difficult to tell if convergence has been achieved
2) Can be wasteful if Markov blanket is large:
\[ P(X_i|MB(X_i)) \] won’t change much (law of large numbers)
Summary

Exact inference by variable elimination:
- polytime on polytrees, NP-hard on general graphs
- space = time, very sensitive to topology

Approximate inference by LW, MCMC:
- LW does poorly when there is lots of (downstream) evidence
- LW, MCMC generally insensitive to topology
- Convergence can be very slow with probabilities close to 1 or 0
- Can handle arbitrary combinations of discrete and continuous variables