Lecture 9
Baysian Networks

Marco Chiarandini

Deptartment of Mathematics & Computer Science
University of Southern Denmark

Slides by Stuart Russell and Peter Norvig

Course Overview

- ✔ Introduction
  - ✔ Artificial Intelligence
  - ✔ Intelligent Agents
- ✔ Search
  - ✔ Uninformed Search
  - ✔ Heuristic Search
- ✔ Adversarial Search
  - ✔ Minimax search
  - ✔ Alpha-beta pruning
- ✔ Knowledge representation and Reasoning
  - ✔ Propositional logic
  - ✔ First order logic
  - ✔ Inference
- Uncertain knowledge and Reasoning
  - Probability and Bayesian approach
  - Bayesian Networks
  - Hidden Markov Chains
  - Kalman Filters
- Learning
  - Decision Trees
  - Maximum Likelihood
  - EM Algorithm
  - Learning Bayesian Networks
  - Neural Networks
  - Support vector machines

Outline

Summary

- Interpretations of probability
- Axioms of Probability
- (Continuous/Discrete) Random Variables
- Prior probability, joint probability, conditional or posterior probability, chain rule
- Inference by enumeration

How to reduce the computation of inference?

Probability basics

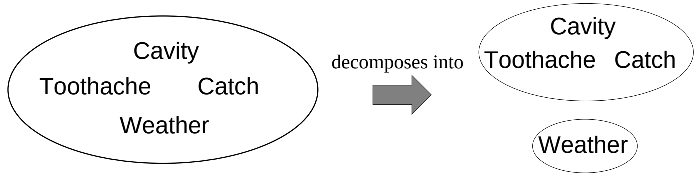
DEFINITION

INDEPENDENT EVENTS
Two events A and B are *independent* of each other if and only if $p(A \cap B) = p(A)\,p(B)$. When $p(B) \neq 0$ this is the same as saying that $p(A) = p(A|B)$. That is, knowing that B is true does not affect the probability of A being true.

CONDITIONALLY INDEPENDENT EVENTS
Two events A and B are said to be *conditionally independent* of each other, given event C if and only if $p((A \cap B) \mid C) = p(A \mid C)\,p(B \mid C)$.

Independence

A and B are independent iff
$\mathbf{P}(A|B) = \mathbf{P}(A)$ or $\mathbf{P}(B|A) = \mathbf{P}(B)$ or $\mathbf{P}(A, B) = \mathbf{P}(A)\mathbf{P}(B)$



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

32 entries reduced to 12; for n independent biased coins, $2^n \to n$

Absolute independence powerful but rare

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

Conditional independence

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

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

$Catch$ is conditionally independent of $Toothache$ given $Cavity$:
$\mathbf{P}(Catch|Toothache, Cavity) = \mathbf{P}(Catch|Cavity)$

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

Conditional independence contd.

Write out full joint distribution using chain rule:
$\mathbf{P}(Toothache, Catch, Cavity)$
$= \mathbf{P}(Toothache|Catch, Cavity)\mathbf{P}(Catch, Cavity)$
$= \mathbf{P}(Toothache|Catch, Cavity)\mathbf{P}(Catch|Cavity)\mathbf{P}(Cavity)$
$= \mathbf{P}(Toothache|Cavity)\mathbf{P}(Catch|Cavity)\mathbf{P}(Cavity)$

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

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 $P(a \wedge b) = P(a|b)P(b) = P(b|a)P(a)$

$$\Longrightarrow \text{Bayes' rule } P(a|b) = \frac{P(b|a)P(a)}{P(b)}$$

or in distribution form

$$\mathbf{P}(Y|X) = \frac{\mathbf{P}(X|Y)\mathbf{P}(Y)}{\mathbf{P}(X)} = \alpha \mathbf{P}(X|Y)\mathbf{P}(Y)$$

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!

Bayes' Rule and conditional independence

$$\begin{aligned} &\mathbf{P}(Cavity|toothache \wedge catch) \\ &= \alpha\, \mathbf{P}(toothache \wedge catch|Cavity)\mathbf{P}(Cavity) \\ &= \alpha\, \mathbf{P}(toothache|Cavity)\mathbf{P}(catch|Cavity)\mathbf{P}(Cavity) \end{aligned}$$

This is an example of a naive Bayes model:

$$\mathbf{P}(Cause, Effect_1, \ldots, Effect_n) = \mathbf{P}(Cause)\prod_i \mathbf{P}(Effect_i|Cause)$$

Total number of parameters is **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

Outline

Outline

- ◇ Syntax
- ◇ Semantics
- ◇ Parameterized distributions

Bayesian networks

Definition

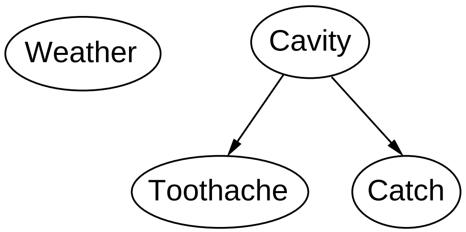
A simple, graphical notation for conditional independence assertions and hence for compact specification of full joint distributions

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) 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*

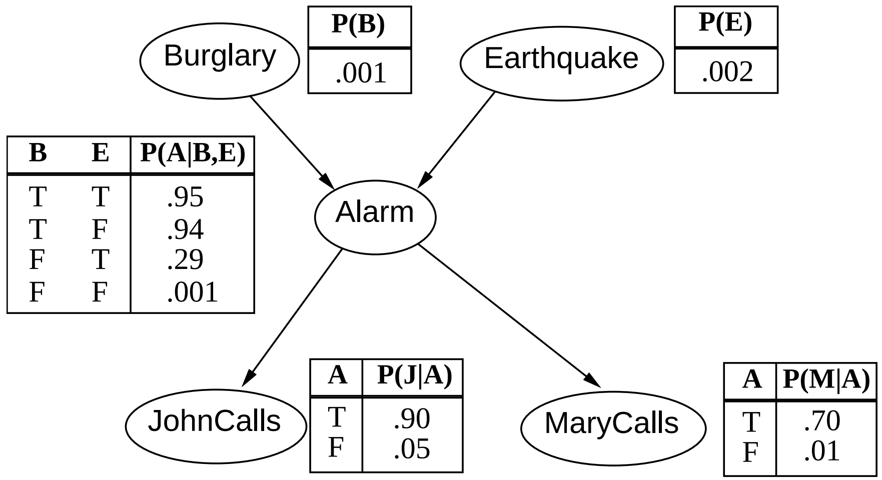
Example

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.



P(B)
.001

P(E)
.002

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

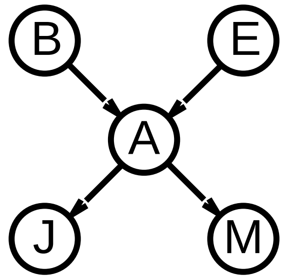
Compactness

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. $2^5 - 1 = 31$)



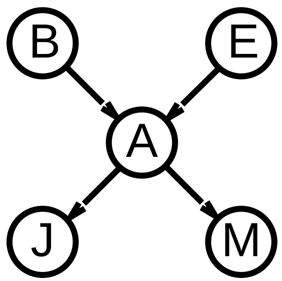
Global semantics

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

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | parents(X_i))$$

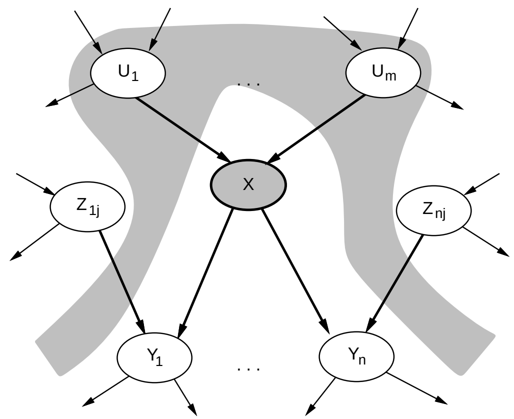
e.g., $P(j \wedge m \wedge a \wedge \neg b \wedge \neg e)$

$$\begin{aligned} &= P(j|a)P(m|a)P(a|\neg b, \neg e)P(\neg b)P(\neg e) \\ &= 0.9 \times 0.7 \times 0.001 \times 0.999 \times 0.998 \\ &\approx 0.00063 \end{aligned}$$



Local semantics

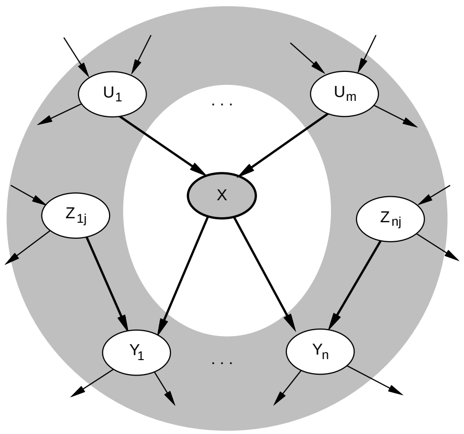
Local semantics: each node is conditionally independent of its nondescendants given its parents



Theorem: Local semantics $\Leftrightarrow$ 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

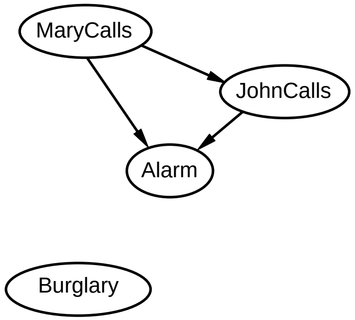
- Choose an ordering of variables $X_1, \ldots, X_n$
- For $i = 1$ to n
  add X_i to the network
  select parents from $X_1, \ldots, X_{i-1}$ such that
  $\mathbf{P}(X_i|Parents(X_i)) = \mathbf{P}(X_i|X_1, \ldots, X_{i-1})$

This choice of parents guarantees the global semantics:

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

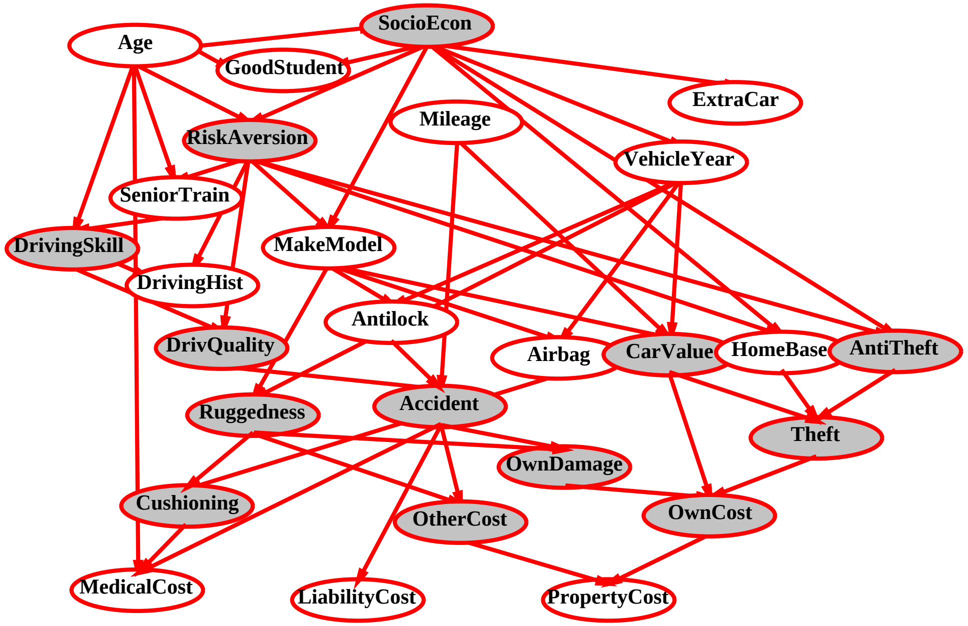
Example

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
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 insurance



Compact conditional distributions

CPT grows exponentially with number 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

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

Compact conditional distributions contd.

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

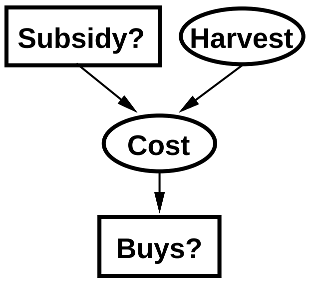
$$\Longrightarrow P(X|U_1 \ldots U_j, \neg U_{j+1} \ldots \neg U_k) = 1 - \prod_{i=1}^{j} q_i$$

$Cold$	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 $Cost$)



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

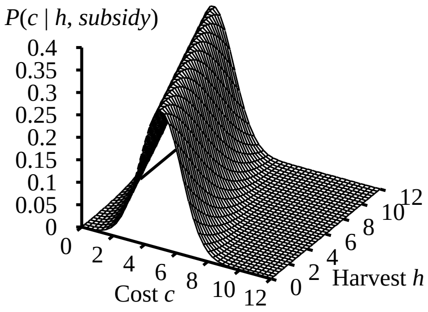
Most common is the linear Gaussian model, e.g.,:

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

Mean $Cost$ varies linearly with $Harvest$, variance is fixed

$\leadsto$Linear variation is unreasonable over the full range
but works OK if the **likely** range of $Harvest$ is narrow

Continuous child variables

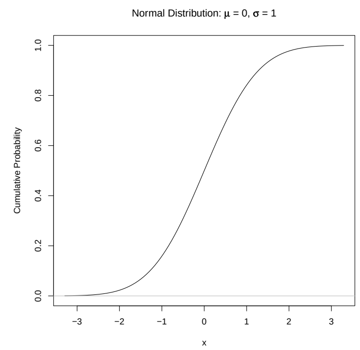


All-continuous network with linear Gaussian distributions
$\Longrightarrow$ full joint distribution is a multivariate Gaussian

Discrete+continuous linear Gaussian 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

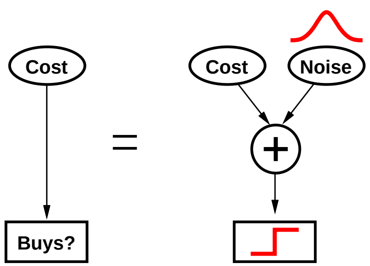
Probability of *Buys?* given *Cost* should be a "soft" threshold:



Probit distribution uses integral of Gaussian:
$\Phi(x) = \int_{-\infty}^{x} N(0,1)(x)dx$
$P(Buys? = true \mid Cost = c) = \Phi((-c + \mu)/\sigma)$

Why the probit?

1. It's sort of the right shape
2. Can be viewed as hard threshold whose location is subject to noise

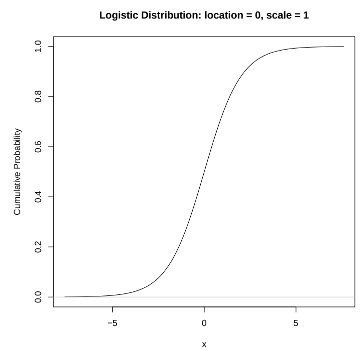


Discrete variable contd.

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:



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 $\Longrightarrow$ parameterized distributions (e.g., linear Gaussian)

Outline

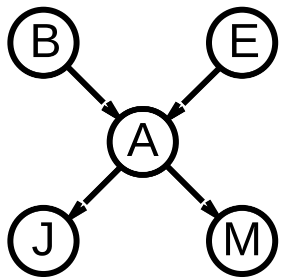
Inference tasks

- Simple queries: compute posterior marginal $\mathbf{P}(X_i | \mathbf{E} = \mathbf{e})$
  e.g., $P(NoGas | Gauge = empty, Lights = on, Starts = false)$
- Conjunctive queries: $\mathbf{P}(X_i, X_j | \mathbf{E} = \mathbf{e}) = \mathbf{P}(X_i | \mathbf{E} = \mathbf{e})\mathbf{P}(X_j | X_i, \mathbf{E} = \mathbf{e})$
- Optimal decisions: decision networks include utility information;
  probabilistic inference required for $P(outcome | action, evidence)$
- 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

Sum out variables from the joint without actually constructing its explicit representation

Simple query on the burglary network:



$$\begin{aligned}\mathbf{P}(B|j,m) &= \mathbf{P}(B,j,m)/P(j,m)\\ &= \alpha\mathbf{P}(B,j,m)\\ &= \alpha \textstyle\sum_e \sum_a \mathbf{P}(B,e,a,j,m)\end{aligned}$$

Rewrite full joint entries using product of CPT entries:

$$\begin{aligned}\mathbf{P}(B|j,m) &= \alpha \textstyle\sum_e \sum_a \mathbf{P}(B)P(e)\mathbf{P}(a|B,e)P(j|a)P(m|a)\\ &= \alpha\mathbf{P}(B) \textstyle\sum_e P(e) \sum_a \mathbf{P}(a|B,e)P(j|a)P(m|a)\end{aligned}$$

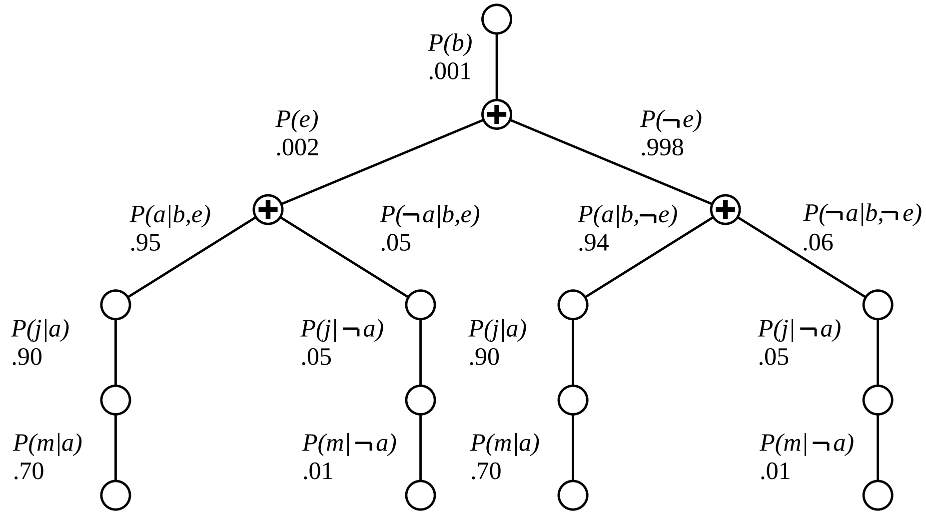
Recursive depth-first enumeration: $O(n)$ space, $O(d^n)$ time

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} ∪ E ∪ Y

   Q(X) ← 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) ← 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 ← First(vars)
   if Y has value y in e
        then return P(y | parent(Y)) × Enumerate-All(Rest(vars), e)
        else return Σ_y P(y | parent(Y)) × Enumerate-All(Rest(vars), e_y)
             where e_y is e extended with Y = y
```


Evaluation tree



Enumeration is inefficient: repeated computation
e.g., computes $P(j|a)P(m|a)$ for each value of e

Complexity of exact inference

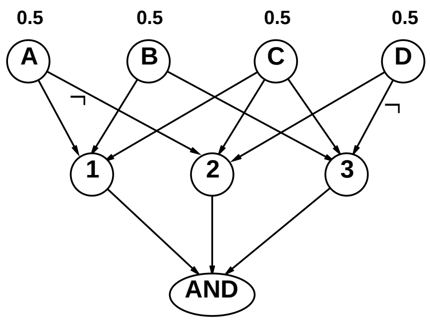
Singly connected networks (or polytrees):
- any two nodes are connected by at most one (undirected) path
- time and space cost (with variable elimination) are $O(d^k n)$
- hence time and space cost are linear in n and k bounded by a constant

Multiply connected networks:
- can reduce 3SAT to exact inference $\Longrightarrow$ NP-hard
- equivalent to **counting** 3SAT models $\Longrightarrow$ #P-complete

1. A v B v C

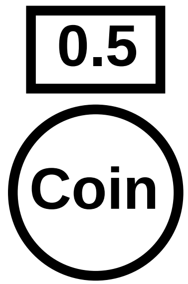
2. C v D v ¬A

3. B v C v ¬D



Inference by stochastic simulation

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



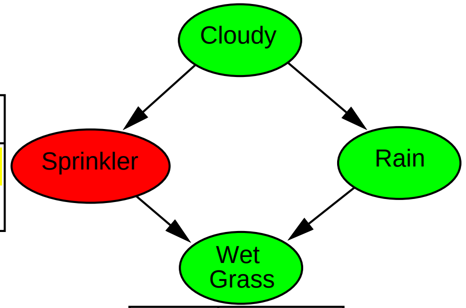
Outline:
- Sampling from an empty network
- Rejection sampling: reject samples disagreeing with evidence
- Likelihood weighting: use evidence to weight samples
- Markov chain Monte Carlo (MCMC): sample from a stochastic process whose stationary distribution is the true posterior

Sampling from an empty network

function Prior-Sample(*bn*) **returns** an event sampled from *bn*
  inputs: *bn*, a belief network specifying joint distribution $\mathbf{P}(X_1, \ldots, X_n)$

  $\mathbf{x} \leftarrow$ an event with n elements
  for $i = 1$ **to** n **do**
    $x_i \leftarrow$ a random sample from $\mathbf{P}(X_i \mid parents(X_i))$
      given the values of $Parents(X_i)$ in $\mathbf{x}$
  return $\mathbf{x}$

Example



P(C)
.50

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

Sampling from an empty network contd.

Probability that PriorSample generates a particular event

$$S_{PS}(x_1 \ldots x_n) = P(x_1 \ldots x_n)$$

i.e., the true prior probability

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

Proof: Let $N_{PS}(x_1 \ldots x_n)$ be the number of samples generated for event $x_1, \ldots, x_n$. Then we have

$$\begin{aligned} \lim_{N\to\infty} \hat{P}(x_1, \ldots, x_n) &= \lim_{N\to\infty} N_{PS}(x_1, \ldots, x_n)/N \\ &= S_{PS}(x_1, \ldots, x_n) \\ &= \prod_{i=1}^{n} P(x_i | parents(X_i)) = P(x_1 \ldots x_n) \end{aligned}$$

$\leadsto$That is, estimates derived from PriorSample are consistent
Shorthand: $\hat{P}(x_1, \ldots, x_n) \approx P(x_1 \ldots x_n)$