DRONACHARYA
College of Engineering

INTELLIGENT SYSTEMS (CSE-303-F)

Section C

Symbolic reasoning under uncertainty

- Problem are not always consistent, complete and unchanging

- **Monotonicity**
  - New facts can be added to the system. If these new facts are consistent with all the other facts that have already been asserted, then nothing will ever be retracted from the set of facts that are known to be true.

Uncertainty

Several Approaches Related to

- Mathematical and Statistical Theories
- Bayesian Statistics
- Fuzzy Sets

Uncertainty in AI

Approximate Reasoning, Inexact Reasoning

Relevant Information is Deficient in One or More

- Information is partial
- Information is not fully reliable
- Representation language is inherently imprecise
- Information comes from multiple sources and it is conflicting
- Information is approximate
- Non-absolute cause-effect relationships exist

- Can include *probability* in the rules
- IF the interest rate is increasing, THEN the price of stocks will decline (80% probability)

Types of Uncertainty

- Uncertainty in prior knowledge

  E.g., some causes of a disease are unknown and are not represented in the background knowledge of a medical-assistant agent

Types of Uncertainty

- Uncertainty in actions

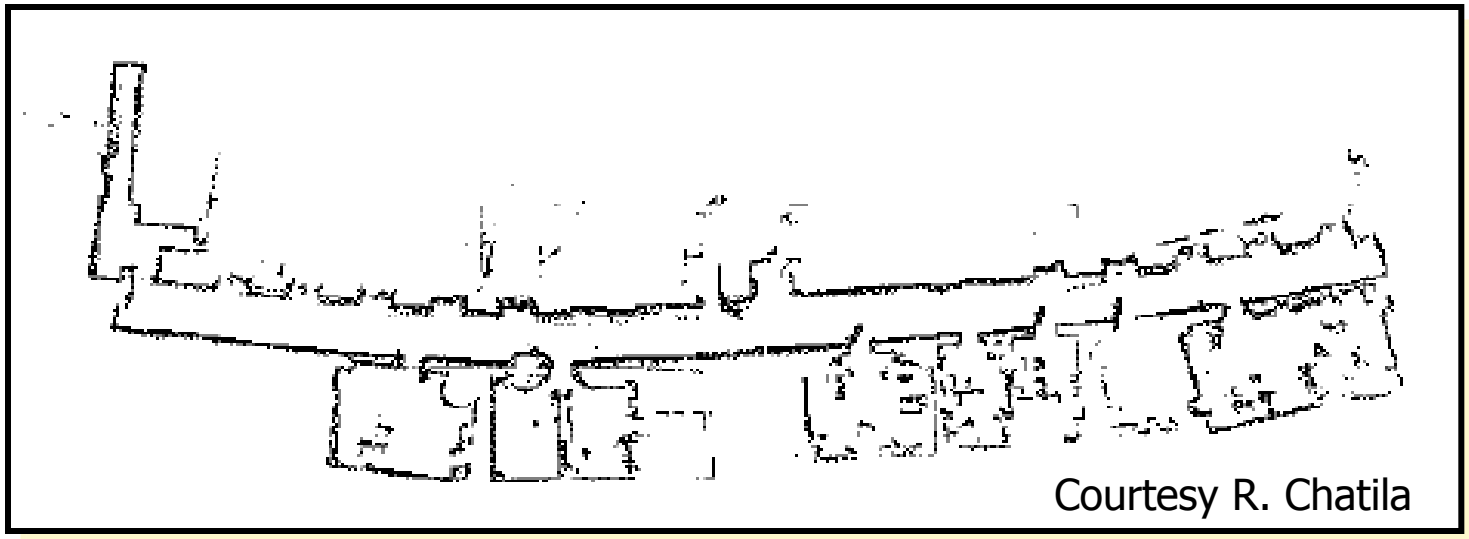
  E.g., to deliver this lecture:
  I must be able to come to college
  my computer must be working
  the LCD projector must be working

As we discussed with planning, actions are represented with relatively short lists of preconditions, while these lists are in fact arbitrary long. It is not efficient (or even possible) to list all the possibilities.

Types of Uncertainty

- Uncertainty in perception

E.g., sensors do not return exact or complete information about the world; a robot never knows exactly its position.



Sources of uncertainty

- Laziness (efficiency)
- Ignorance

What we call *uncertainty* is a summary of all that is not explicitly taken into account
in the agent's knowledge base (KB).

Assumptions of reasoning with predicate logic

(1). Predicate descriptions must be sufficient with respect to the application domain.

Each fact is known to be either true or false. But what does lack of information mean?

- Closed world assumption, assumption based reasoning:

  PROLOG: if a fact cannot be proven to be true, assume that it is false

  HUMAN: if a fact cannot be proven to be false, assume it is true

Assumptions of reasoning with predicate logic (cont'd)

(2). The information base must be consistent.

- Human reasoning: keep alternative (possibly conflicting) hypotheses. Eliminate as new evidence comes in.

Assumptions of reasoning with predicate logic (cont'd)

(3). Known information grows *monotonically* through the use of inference rules.

- Need mechanisms to:
  - add information based on assumptions (nonmonotonic reasoning), and
  - delete inferences based on these assumptions in case later evidence shows that the assumption was incorrect (truth maintenance).

Questions

- How to represent uncertainty in knowledge?
- How to perform inferences with uncertain knowledge?
- Which action to choose under uncertainty?

Approaches to handling uncertainty

- Default reasoning [Optimistic]
  non-monotonic logic
- Worst-case reasoning [Pessimistic]
  adversarial search:
  - *Adversarial search, or game-tree search, is a technique for analyzing an adversarial game in*

  order to try to determine who can win the game and what moves the players should make in order to win.
- Probabilistic reasoning [Realist]
  probability theory

Default Reasoning

- Rationale: The world is fairly normal. Abnormalities are rare.
- So, an agent assumes normality, until there is evidence of the contrary.
- E.g., if an agent sees a bird X, it assumes that X can fly, unless it has evidence that X is a penguin, an ostrich, a dead bird, a bird with broken wings, …

Modifying logic to support nonmonotonic inference

- $p(X) \wedge$ unless $q(X) \rightarrow r(X)$

- If we
- believe p(X) is true, and
- do not believe q(X) is true (either unknown or believed to be false)
- then we
- can infer r(X)
- later if we find out that q(X) is true, r(X) must be retracted

"unless" is a *modal operator*: deals with belief rather than truth

Modifying logic to support nonmonotonic inference (cont'd)

- $p(X) \wedge$ unless $q(X) \rightarrow r(X)$ in KB
- $p(Z)$ in KB
- $r(W) \rightarrow s(W)$ in KB
- - - - - - -
- $\neg q(X)$?? $q(X)$ is not in KB
- $r(X)$ inferred
- $s(X)$ inferred

Example

- If it is snowing and unless there is an exam tomorrow, I can go skiing.
- It is snowing.
- Whenever I go sking, I stop by at the Chalet to drink hot chocolate.
- - - - - - -
- I did not check my calendar but I don’t remember an exam scheduled for tomorrow, conclude: I’ll go skiing. Then conclude: I’ll drink hot chocolate.

"Abnormality"

- $p(X) \wedge$ unless ab $p(X) \rightarrow q(X)$
- ab: abnormal
- Examples: If X is a bird, it will fly unless it is abnormal.
  - (abnormal: broken wing, sick, trapped, ostrich, ...)
- If X is a car, it will run unless it is abnormal.
  - (abnormal: flat tire, broken engine, no gas, …)

Another modal operator: M

- $p(X) \wedge M\ q(X) \rightarrow r(X)$
- If
  - we believe p(X) is true, and
  - q(X) is *consistent with* everything else,
- then we
  - can infer r(X)

"M" is a *modal operator* for "is consistent."

Probabilities and Related Approaches

The Probability Ratio

- P(X) = Number of outcomes favoring the occurrence of X

Total number of outcomes

- Multiple Probability Values in Many Systems
  - Three-part antecedent (probabilities: 0.9, 0.7, and 0.65)
  - The overall probability:

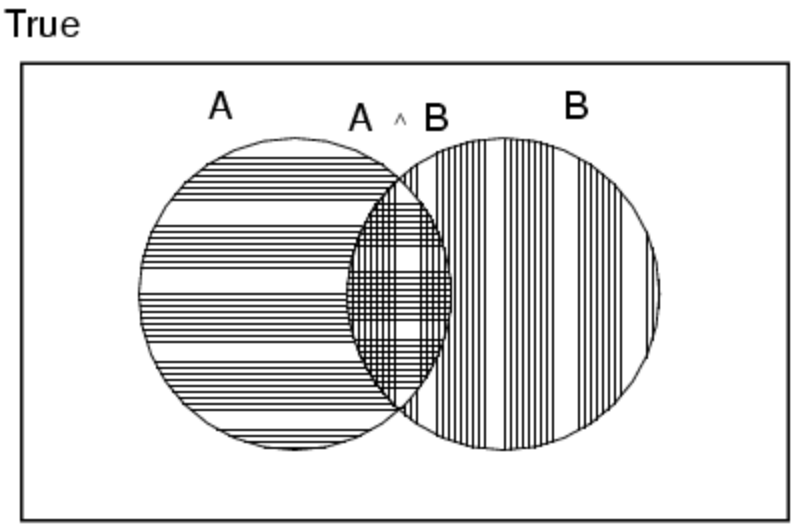
    P = (0.9)(0.7)(0.65) = 0.4095
- Sometimes one rule references another - individual rule probabilities can propagate from one to another

Several Approaches for Combining Probabilities

- Probabilities can be
  - **Multiplied (joint probabilities)**
  - **Averaged (simple or a weighted average)**
  - **Highest value**
  - **Lowest value**
- Rules and events are considered independent of each other
- If Dependent - Use the Bayes extension theorem

Axioms of probability

- For any propositions *A, B*
-
  - $0 \leq P(A) \leq 1$
  - $P(true) = 1$ and $P(false) = 0$
  - $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$
  -



BAYESIAN APPROACH

- It depend on the use of known prior and likely probabilities.
- It is introduced by Clergyman Thomas Bayes in the eighteenth Century.
- It depends on the use of conditional probabilities of specified events when it is known that other events have occurred.

- For two events H and E with the Probability P(E) > 0, the conditional probability of event H, given that event E has occurred , is defined as
  - P (H/E) = P(H & E) /P(E) 1
- The Conditional probability of event E given that event H occurred can likewise be written
  - P (E/H) = P(H & E) /P(H) 2
- From 1 & 2
  - P (H/E)=P(E/H)P/(H) /P(E) 3

EXAMPLE

- Patient has the certain disease D1 given the symptom E we wish to find out P(D1/E).
  - P(D1) =0.05 and P(E) = 0.15 P(E/D1) =0.95
  - P(D1/E) = 0.32

- If P(E) is difficult to obtain , then replace H with ~H in equation 3.
- P (~H/E)=P(E/~H)P/(~H) /P(E) 4

- Divide equation ¾
- P (H/E)/ P (~H/E)= P(E/H)P/(H) / P(E/~H)P/(~H)
- O (H/E) = L(E/H).O(H) 5
- Posterior odds = likelihood ratio * prior odds on H.

Syntax

- Basic element: random variable
- Similar to propositional logic: possible worlds defined by assignment of values to random variables.
- Boolean random variables
- e.g., *Cavity* (do I have a cavity?)
- Discrete random variables
- e.g., *Weather* is one of <*sunny,rainy,cloudy,snow*>
- Domain values must be exhaustive and mutually exclusive

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

Suppose I believe the following:

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

Syntax

- Atomic event: A complete specification of the state of the world about which the agent is uncertain

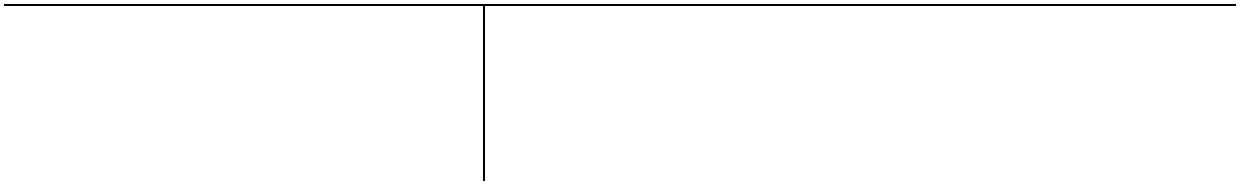
  E.g., if the world consists of only two Boolean variables *Cavity* and *Toothache*, then there are 4 distinct atomic events:

  Cavity = false $\wedge$ *Toothache = false*
  Cavity = false $\wedge$ *Toothache = true*
  Cavity = true $\wedge$ *Toothache = false*
  Cavity = true $\wedge$ *Toothache = true*

  Atomic events are mutually exclusive and exhaustive

Prior probability

- 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*) = <0.72,0.1,0.08,0.1> (normalized, i.e., sums to 1)
- Joint probability distribution for a set of random variables gives the probability of every atomic event on those random variables



P(*Weather,Cavity*) = a 4 × 2 matrix of values:

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

- Every question about a domain can be answered by the joint distribution

Conditional probability

- Conditional or posterior probabilities
- e.g., P(*cavity* | *toothache*) = 0.8

  i.e., given that *toothache* is all I know

- (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(*cavity* | *toothache,cavity*) = 1

- New evidence may be irrelevant, allowing simplification, e.g.,
- P(*cavity* | *toothache, sunny*) = P(*cavity* | *toothache*) = 0.8
- This kind of inference, sanctioned by domain knowledge, is crucial

Conditional probability

- Definition of conditional probability:
- $P(a \mid b) = P(a \wedge b) / P(b)$ if $P(b) > 0$
- Product rule gives an alternative formulation:

  $P(a \wedge b) = P(a \mid b)\, P(b) = P(b \mid a)\, P(a)$

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

P(*Weather,Cavity*) = P(*Weather* | *Cavity*) P(*Cavity*)
(View as a set of 4 × 2 equations, not matrix mult.)

- Chain rule is derived by successive application of product rule:
- $P(X_1, \ldots, X_n) = P(X_1,\ldots,X_{n-1})\, P(X_n \mid X_1,\ldots,X_{n-1})$
  $= P(X_1,\ldots,X_{n-2})\, P(X_{n-1} \mid X_1,\ldots,X_{n-2})\, P(X_n \mid X_1,\ldots,X_{n-1})$
  $= \ldots$
  $= \pi_{i=1}^{n} P(X_i \mid X_1, \ldots, X_{i-1})$

Inference by enumeration

- Start with the joint probability 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: $P(\varphi) = \Sigma_{\omega:\omega \models \varphi} P(\omega)$

Inference by enumeration

- Start with the joint probability 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: $P(\varphi) = \Sigma_{\omega:\omega \models \varphi} P(\omega)$
- P(*toothache*) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2

Inference by enumeration

- Start with the joint probability 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: $P(\varphi) = \Sigma_{\omega:\omega \models \varphi} P(\omega)$
- P(*toothache*) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2

Inference by enumeration

- Start with the joint probability distribution:

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

- Can also compute conditional probabilities:

$$P(\neg cavity \mid toothache) = \frac{P(\neg cavity \wedge toothache)}{P(toothache)}$$

$$= \frac{0.016+0.064}{0.108 + 0.012 + 0.016 + 0.064}$$

$$= 0.4$$

Normalization

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

- Denominator can be viewed as a normalization constant α
- P(*Cavity* | *toothache*) = α, P(*Cavity,toothache*)
  = α, [P(*Cavity,toothache,catch*) + P(*Cavity,toothache*,¬ *catch*)]
  = α, [<0.108,0.016> + <0.012,0.064>]
  = α, <0.12,0.08> = <0.6,0.4>

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:

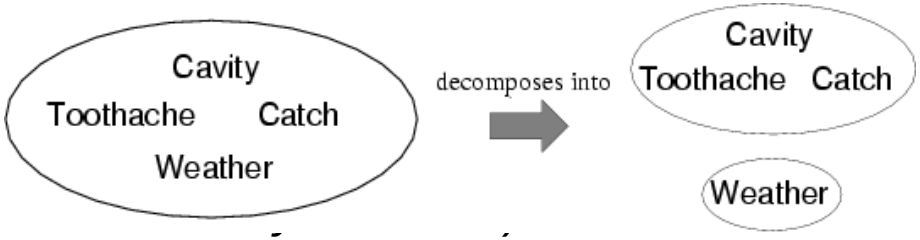
$$P(Y \mid E = e) = \alpha P(Y, E = e) = \alpha \Sigma_h P(Y, E = e, H = h)$$

- 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.
  2. Space complexity $O(d^n)$ to store the joint distribution
  3.
  3. How to find the numbers for $O(d^n)$ entries?

INDEPENDENCE

- *A* and *B* are independent iff
  P(*A*|*B*) = P(*A*) or P(*B*|*A*) = P(*B*) or P(A, B) = P(*A*) P(*B*)



P(*Toothache,*
  = P(*Toothache, Catch, Cavity*) P(*Weather*)

- 32 entries reduced to 12; for *n* independent biased coins, $O(2^n) \rightarrow O(n)$
- 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) P(*catch* | *toothache, cavity*) = P(*catch* | *cavity*)

- The same independence holds if I haven't got a cavity:
- (2) P(*catch* | *toothache*,¬*cavity*) = P(*catch* | ¬*cavity*)

- *Catch* is conditionally independent of *Toothache* given *Cavity*:
- P(*Catch* | *Toothache,Cavity*) = P(*Catch* | *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

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.

- Equivalent statements:

  P(*Toothache* | *Catch, Cavity*) = P(*Toothache* | *Cavity*)

  P(*Toothache, Catch* | *Cavity*) = P(*Toothache* | *Cavity*) P(*Catch* | *Cavity*)

Bayes' Rule

- Product rule $P(a \wedge b) = P(a \mid b)\, P(b) = P(b \mid a)\, P(a)$
  $\Rightarrow$ Bayes' rule: $P(a \mid b) = P(b \mid a)\, P(a) / P(b)$

- or in distribution form
- $P(Y|X) = P(X|Y)\, P(Y) / P(X) = \alpha P(X|Y)\, P(Y)$

- Useful for assessing diagnostic probability from causal probability:
  - P(Cause|Effect) = P(Effect|Cause) P(Cause) / P(Effect)
  - E.g., let *M* be meningitis, *S* be stiff neck:
  - P(m|s) = P(s|m) P(m) / P(s) = 0.8 × 0.0001 / 0.1 = 0.0008
  - Note: posterior probability of meningitis still very small!

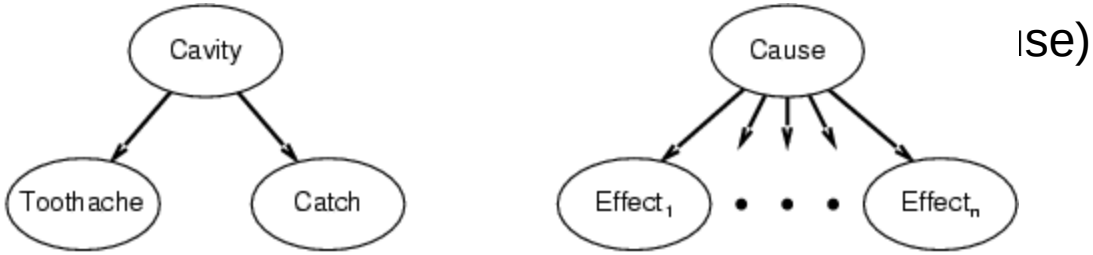
Bayes' Rule and conditional independence

P(*Cavity* | *toothache* $\wedge$ *catch*)

= αP(*toothache* $\wedge$ *catch* | *Cavity*) P(*Cavity*)

= αP(*toothache* | *Cavity*) P(*catch* | *Cavity*) P(*Cavity*)

- This is an example of a naïve Bayes model:
- P(Caus



se)

- 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

DEMPSTER-SHAFER THEORY

- Drawbacks of using Bayesian theory
  - The probabilities are described as a single numeric point value.
    - Distortion to precision that is actually available for supporting evidence.
    - When we assert with probability 0.7 that the dollar will fall against the Japensese Yen over the next six months, what we really mean is we have a fairly strong conviction there is a chance of about 0.6 to 0.8 say, that it will fall

DEMPSTER-SHAFER THEORY

- Drawbacks of using Bayesian theory
  - No way to differentiate between ignorance and uncertainty.
  - Example
    - One of the three A, B, C terrorist group has planted a bomb. Let C found guilty and P(C) = 0.8. According to traditional theory rest of the probability will distributes amongst other without having any knowledege about them.

DEMPSTER-SHAFER THEORY

- Drawbacks of using Bayesian theory
  - Forced to regard belief and disbelief as functional opposite.
    - Ex. If P(A) = 0.3 then P(~A)= 0.7 so that P(A) + P(~A) =1

AS A REMEDY FOR THE ABOVE PROBLEMS, GENERALISES THEORY HAS BEEN PROPOSED BY ARTHUR DEMPSTER (1968) AND EXTENDED BY STUDENT GLENN SHAFER (1976).

DEMPSTER-SHAFER THEORY

- Separate probability masses may be assigned to all subsets of a universe of discourse rather than just to individual single members.
  - It Permit the inequality P(A) + P(~A) <=1
    - It assume a universe of discourse U
    - A set corresponding to n proposition, exactly one of which is true.
    - The Propositions are assumed to be exhaustive and mutually exclusive.

DEMPSTER-SHAFER THEORY

- Let 2^u denote all subsets of U.
- Let the set function m defined on 2^u , be a mapping to [0,1],
  - m : 2^u → [0,1], be, such that for all subsets A C U
  - m ($) =0
  - $\sum_{A \subseteq U} m(A) = 1$

- The function m defines a probability distribution on 2^u .(not just on the singletons of U as in classical theory)
- It represent the measure of belief committed exactly to A.
- IT IS POSSIBLE TO ASSIGN BELIEF TO EACH SUBSET A OF U WITHOUT ASSIGNING ANY TO ANYTHING SMALLER.

DEMPSTER-SHAFER THEORY

- $Bel(A) = \sum_{B \subseteq A} m(B)$

A BELIEF FUNCTION, BEL , CORRESPONDING TO A SPECIFIC m FOR THE SET A, IS DEFINED AS THE SUM OF BELIEFS COMMITTED TO EVERY SUBSET OA A BY m.

Ex: if U contain the mutually exclusive subsets A, B, C and D then

Bel({A,C,D})=m({A,C,D}) + m({A,C}) + m({A,D}) + m({C,D}) + m({A}) + m({C}) + m({D})

DEMPSTER-SHAFER THEORY

- Some related terms and facts:
  - In D-S T, a belief interval can also be defined for a subset A. It is represented as the subinterval [Bel(A), P1(A)] of [0,1].
  - Support: Bel(A) is also called support of A:
  - Plausibility: P1(A) =1-Bel(~A), the plausibility of A.
  - Focal Element: The subsets A of U are called the focal elements of the support function Bel when m(A)>0.

DEMPSTER-SHAFER THEORY

- Some related terms and facts:
  - Doubts: Bel(A) partially describes the beliefs about proposition A, belief in ~ A (doubt) can be defined as D(A)= Bel(~A)
  - Hence Upper bound of the interval can be defined as P1(A) = 1- D(A) =1-Bel(~A).
  - Confidence: The Belief interval,[Bel(A), P1(A)], is also sometimes refereed to as a confidence in A, while the quantity P1(A) –Bel(A) is referred to as the uncertainty in A.

DEMPSTER-SHAFER THEORY

- It can be shown that (Prade 1983)
  - P1($) =0, P1(U)=1
  - For all A
    - P1(A) >=Bel(A),
    - Bel(A)+Bel(~A)<=1,
    - P1(A) +P1(~A)>=1, and
  - For ACB,
    - Bel(A) <=Bel(B), P1(A)<=P1(B)
- In interpreting the above definitions, it should be noted that a portion of belief may be committed to a set of propositions, but need not be, and if committed, it is not necessary to its negation. However, a belief committed to a proposition is committed to any other proposition it implies.

DEMPSTER-SHAFER THEORY

- [0,1] represents no belief in support of the proposition
- [0,0] represents the belief the proposition is false
- [1,1] represents the belief the proposition is true
- [.3,1] represents partial belief in the proposition
- [0,.8] represents partial disbelief in the proposition
- [.2,.7] represents belief from evidence both for and against the proposition

DEMPSTER-SHAFER THEORY

- When evidence is available from two or more independent KS Bel1 and Bel2 then Demster has provided such a combining function denoted by Bel1 o Bel2.
  - Let m1 and m2 are two basic probabbility assignment function to the Bel1 and Bel2
  - Let A1, ...Ak be focal elements for Bel1 and B1, ..., Bp be the focal elements for Bel2.
  - m1(Ai) and m2(Bj) each assign probability masses on the unit interval. They can be orthogonally represented as:

DEMPSTER-SHAFER THEORY

m1(A1) . . .. .m1(Ai). . . . ..

.

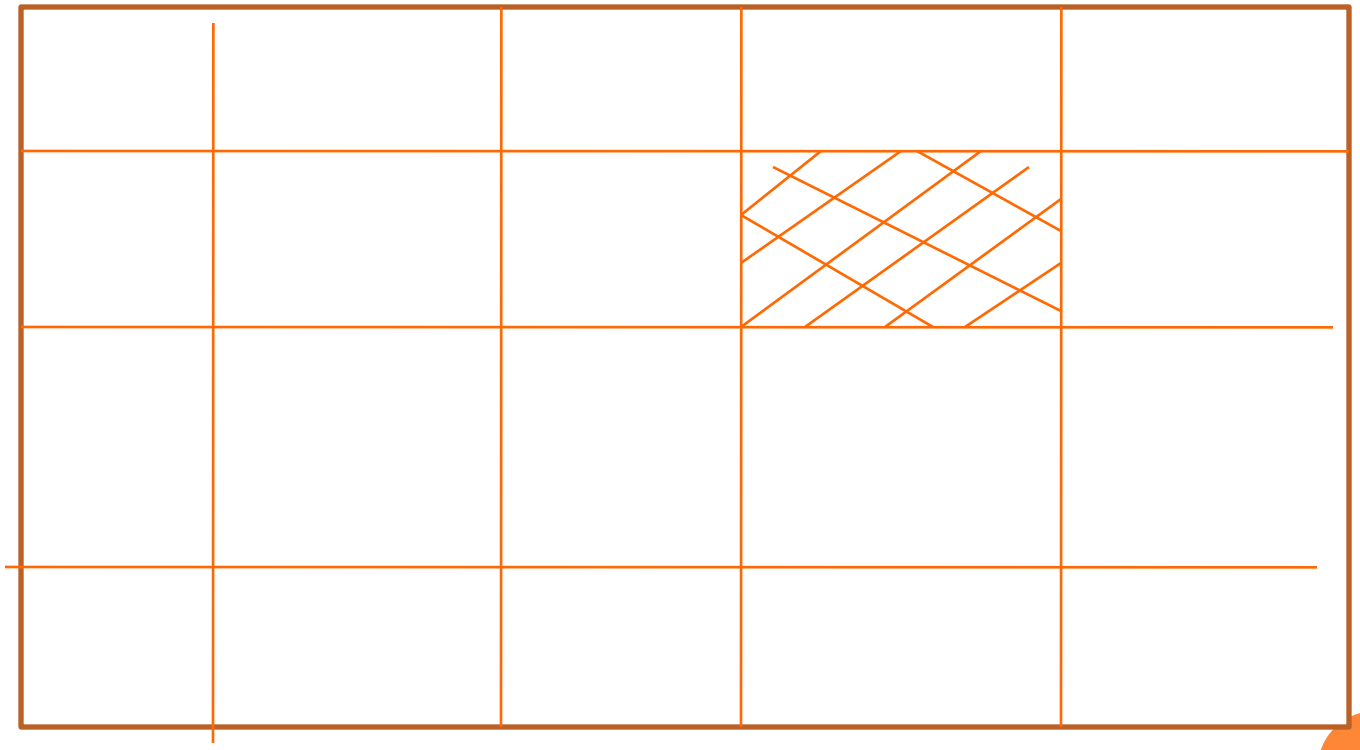
.

m2(Bj)

.

.

m2(B1)



DEMPSTER-SHAFER THEORY

- The unit square square represents the total probability mass assigned by both m1 and m2 for all of their common subsets.
- Particular rectangle within the square, shown as the intersection of the sets Ai and Bj, has committed to it the measure m1(Ai)m2(Bj). Therefore the total probability mass committed to C will be
  - ∑m1(Ai)m2(Bj) Its normalised form after removing

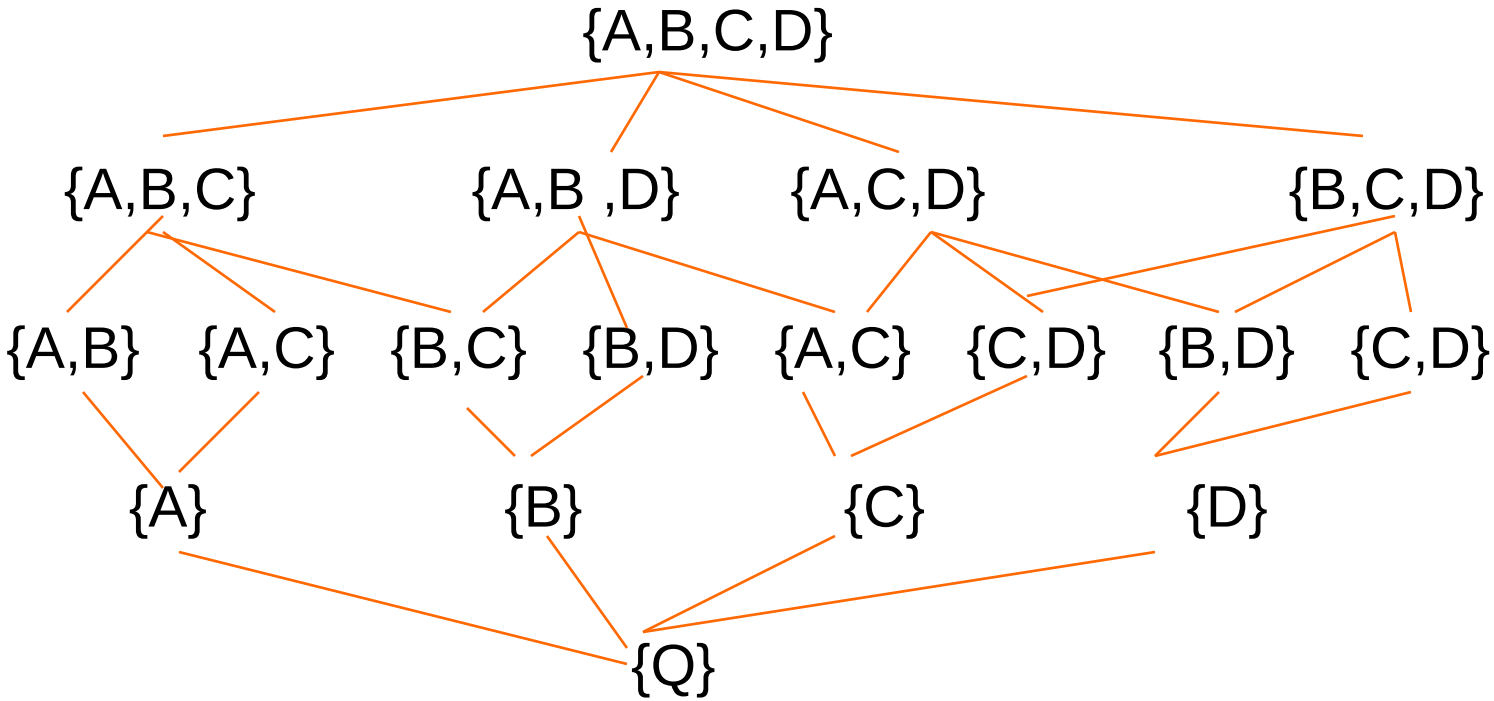
Ai^Bj=c

Ai^Bi=@

$$m1 \text{ o } m2 = \sum_{Ai \wedge Bj} m1(Ai)m2(Bj) \Big/ \sum_{Ai \wedge Bj \neq 0} m1(Ai)m2(Bj$$

DEMPSTER-SHAFER THEORY

- Ex of problem of identifying the terrorist group from A, B, C, D THE POSSIBLE CASES ARE



DEMPSTER-SHAFER THEORY

- Assume that one group A and C were responsible to a degree of m1({A,C}) =0.6,
- ANOTHER SOURCE OF EVIDENCE DISPROVES THE BELIEF THAT c WAS INVOLVED hence it means m2({A,B,D}) = 0.7.
- To obtain the pooled evidence, we compute the following quantitites.
  - m1 o m2({A}) = 0.6 * 0.7 = 0.42
  - m1 o m2({A,C}) = 0.6 * 0.3 = 0.18
  - m1 o m2({A,B,D}) = 0.4 * 0.7 = 0.28
  - m1 o m2({U}) = 0.4 * 0.3 =0.12
  - m1 o m2 for all other set =0
  - Bel1({A,C}) = m({A,C}) + m({A}) + m({C})

Bayesian networks

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

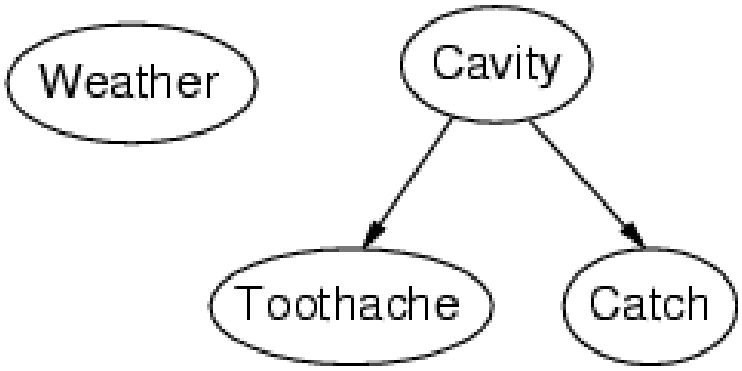
- Syntax:
  - a set of nodes, one per variable
  -
  - a directed, acyclic graph (link ≈ "directly influences")
  - a conditional distribution for each node given its parents:

$$\mathbf{P}(X_i \mid \text{Parents}(X_i))$$

- 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*

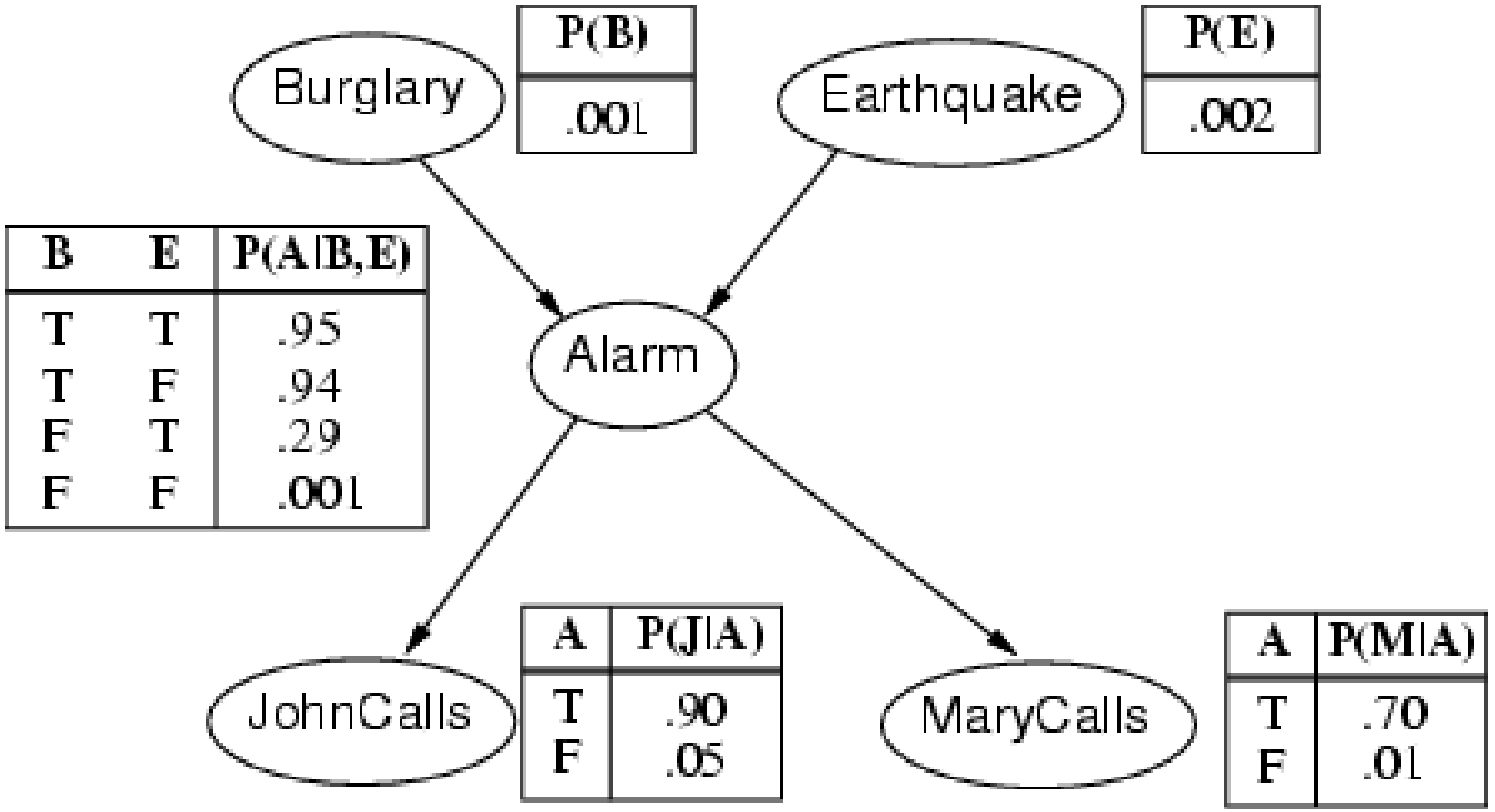
Example

- 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: *Burglary*, *Earthquake*, *Alarm*, *JohnCalls*, *MaryCalls*

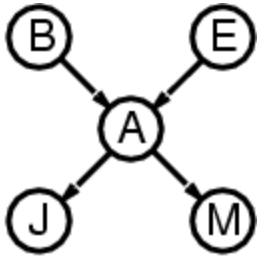
- Network topology reflects "causal" knowledge:
  - A burglar can set the alarm off
  - An earthquake can set the alarm off
  - The alarm can cause Mary to call
  - The alarm can cause John to call

EXAMPLE CONTD.



Compactness

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

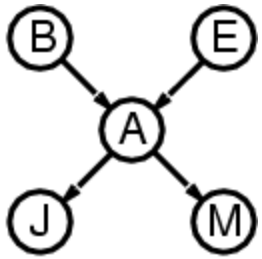


- Each row requires one number p for X_i = *true* (the number for X_i = *false* is just *1-p*)
- If each variable has no more than k parents, the complete network requires $O(n \cdot 2^k)$ numbers
- I.e., grows linearly with n, vs. $O(2^n)$ for the full joint distribution
- For burglary net, 1 + 1 + 4 + 2 + 2 = 10 numbers (vs. $2^5-1 = 31$)

Semantics

The full joint distribution is defined as the product of the local conditional distributions:

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



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

$= P(j \mid a)\, P(m \mid a)\, P(a \mid \neg b, \neg e)\, P(\neg b)\, P(\neg e)$

Constructing Bayesian networks

- 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

$$P(X_i \mid Parents(X_i)) = P(X_i \mid X_1, \ldots X_{i-1})$$

This choice of parents guarantees:

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

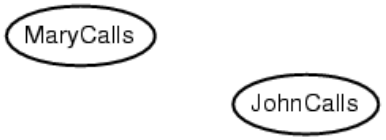
(chain rule)

$$= \prod_{i=1}^{n} P(X_i \mid Parents(X_i))$$

(by construction)

Example

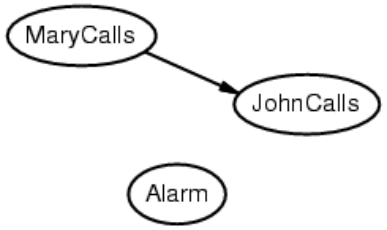
- Suppose we choose the ordering *M, J, A, B, E*
-



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

Example

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



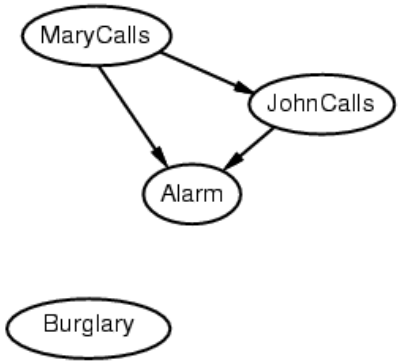
P(J | M) = P(J)?

No

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

Example

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



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

No

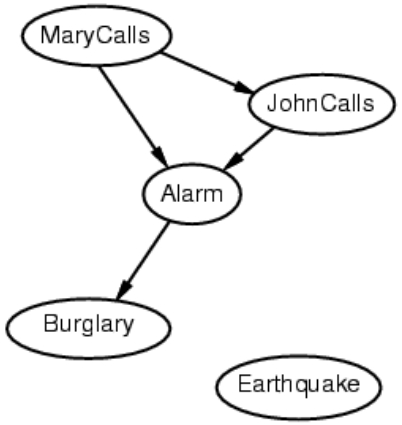
P(A | J, M) = P(A | J)? P(A | J, M) = P(A)? No

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

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

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

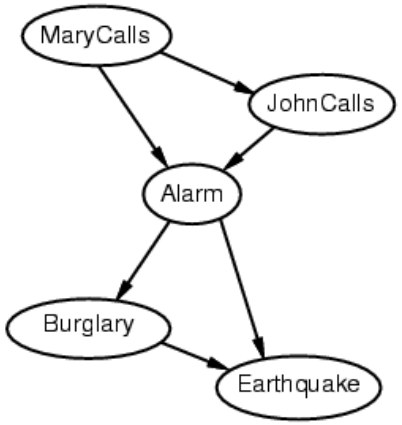
P(B | A, J, M) = P(B | A)? Yes

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

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

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

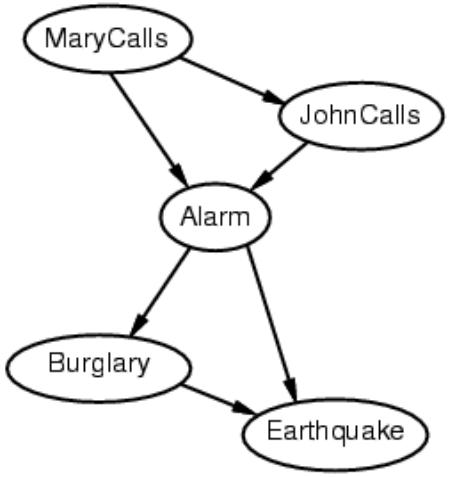
P(B | A, J, M) = P(B | A)? Yes

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

$P(E \mid B, A, J, M) = P(E \mid A)$? No

$P(E \mid B, A, J, M) = P(E \mid A, B)$? Yes

Example contd.

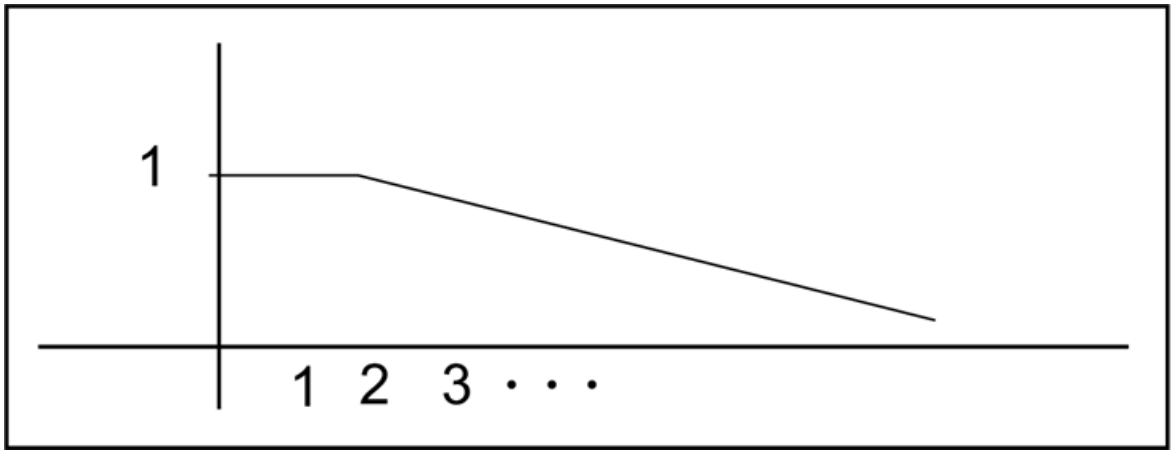


- Deciding conditional independence is hard in noncausal directions
- (Causal models and conditional independence seem hardwired for humans!)
- Network is less compact: 1 + 2 + 4 + 2 + 4 = 13 numbers needed
-

Summary

- Bayesian networks provide a natural representation for (causally induced) conditional independence
- Topology + CPTs = compact representation of joint distribution
- Generally easy for domain experts to construct

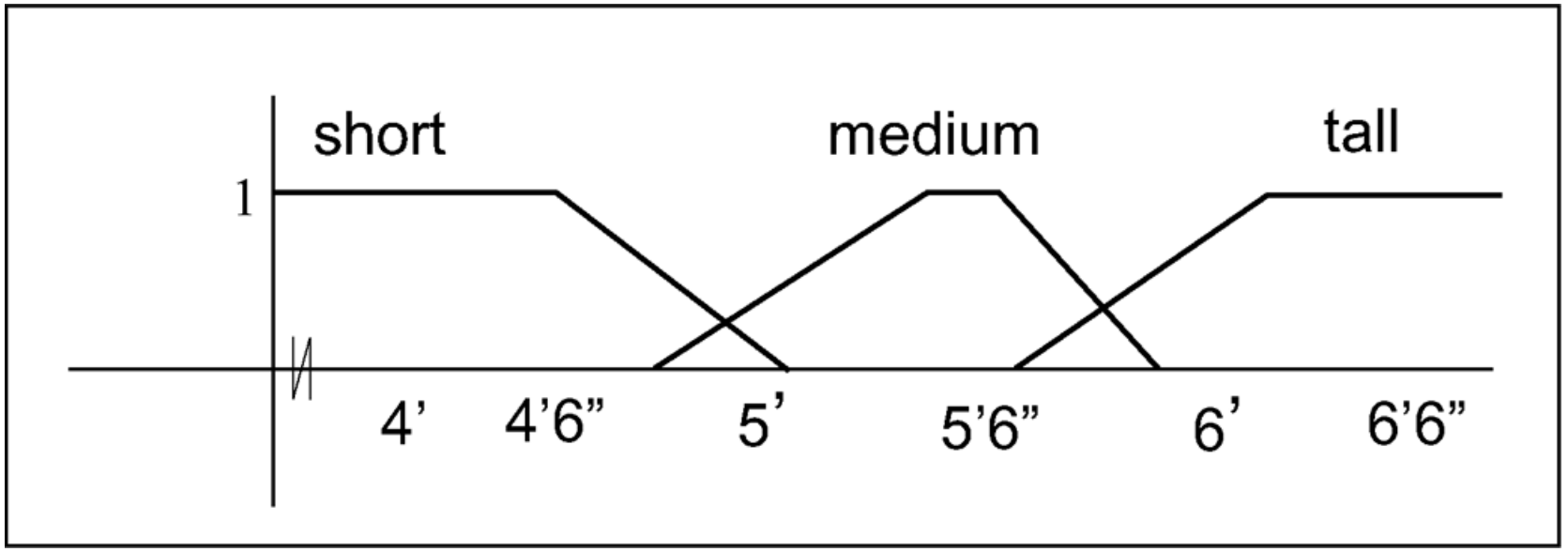
The fuzzy set representation for "small integers"



Reasoning with fuzzy sets

- Lotfi Zadeh's fuzzy set theory
- Violates two basic assumption of set theory
  - For a set S, an element of the universe either belongs to S or the complement of S.
  - For a set S, and element cannot belong to S or the complement S at the same time
- Jack is 5'7". Is he tall? Does he belong to the set of tall people? Does he not belong to the set of tall people?

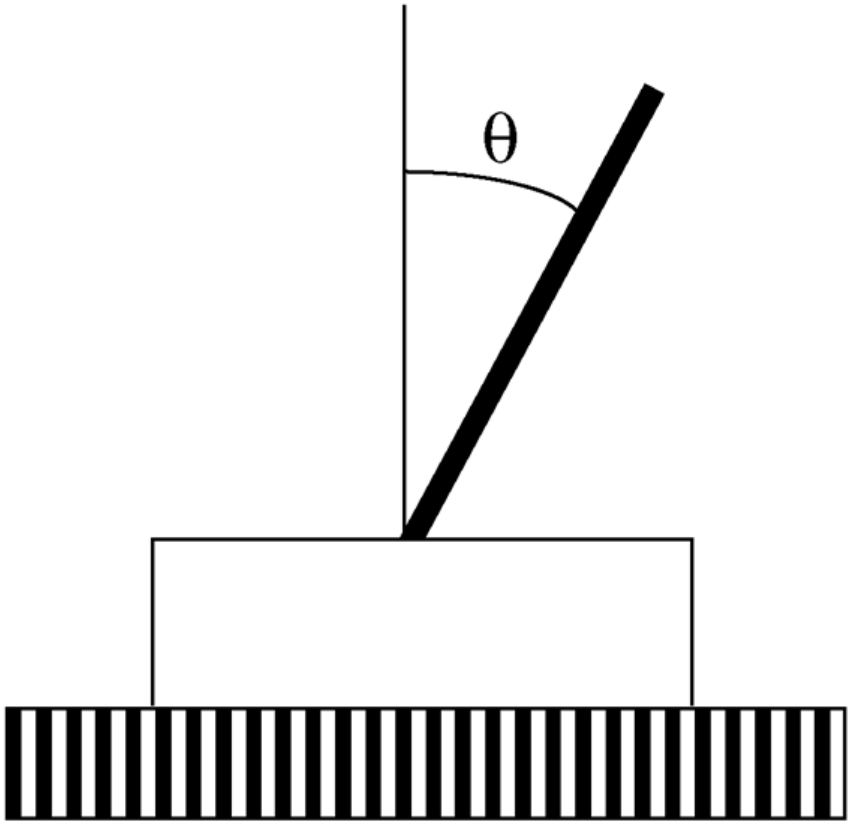
A fuzzy set representation for the sets short, median, and tall males



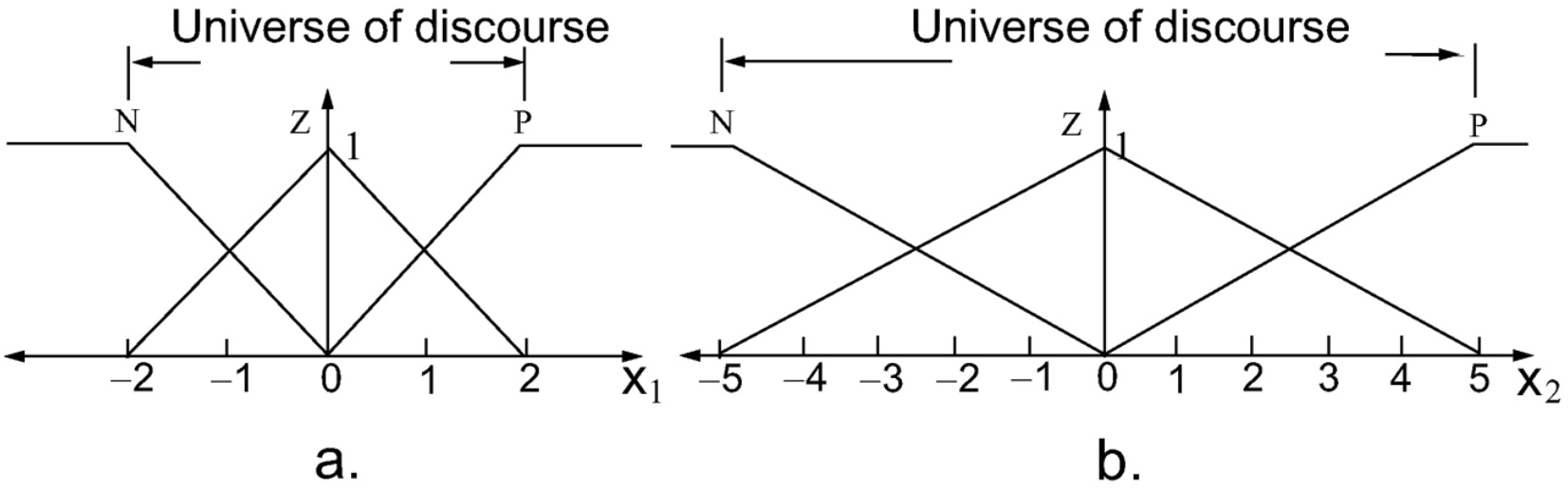
Fuzzy logic

- Provides rules about evaluating a fuzzy truth, T
- The rules are:
  - $T(A \wedge B) = \min(T(A), T(B))$
  - $T(A \vee B) = \max(T(A), T(B))$
  - $T(\neg A) = 1 - T(A)$
- Note that unlike logic $T(A \vee \neg A) \neq T(True)$

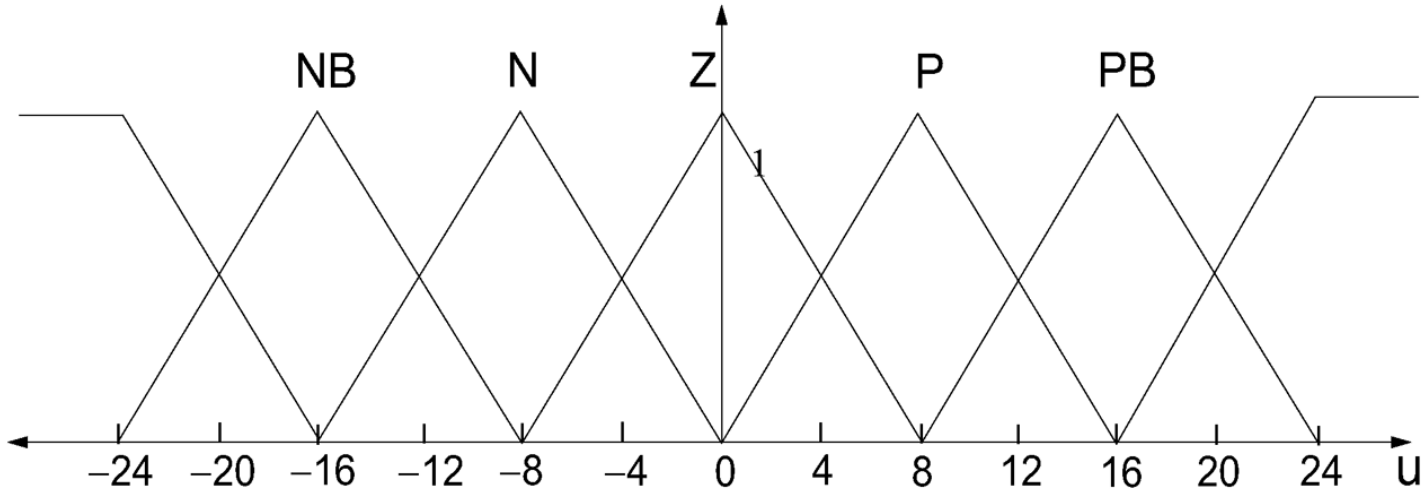
The inverted pendulum and the angle θ and dθ/dt input values.



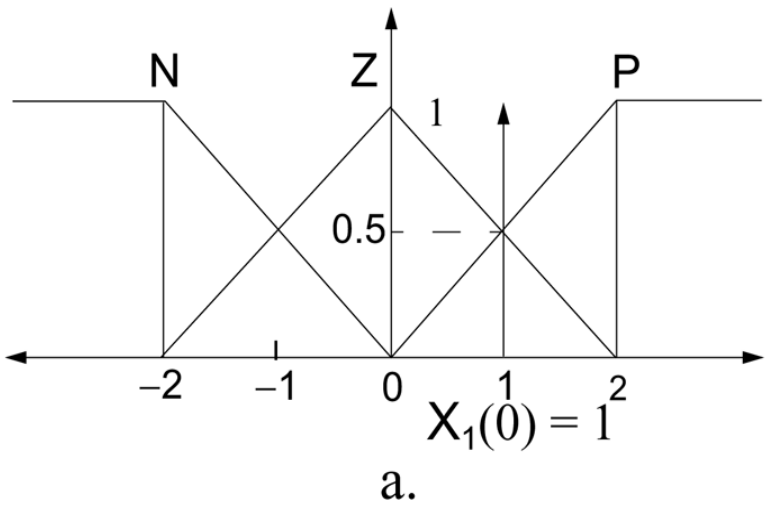
The fuzzy regions for the input values θ (a) and dθ/dt (b)



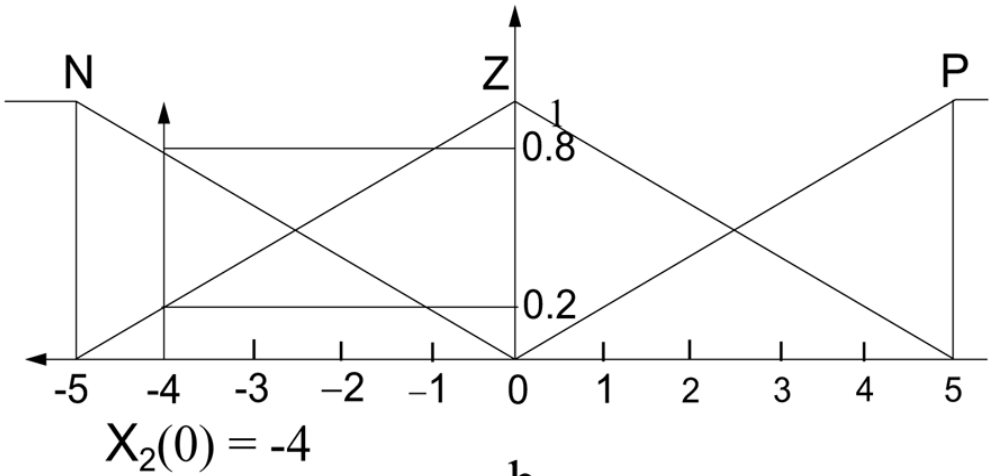
The fuzzy regions of the output value U, indicating the movement of the pendulum base



The fuzzification of the input measures x1=1, x2 = -4



a.

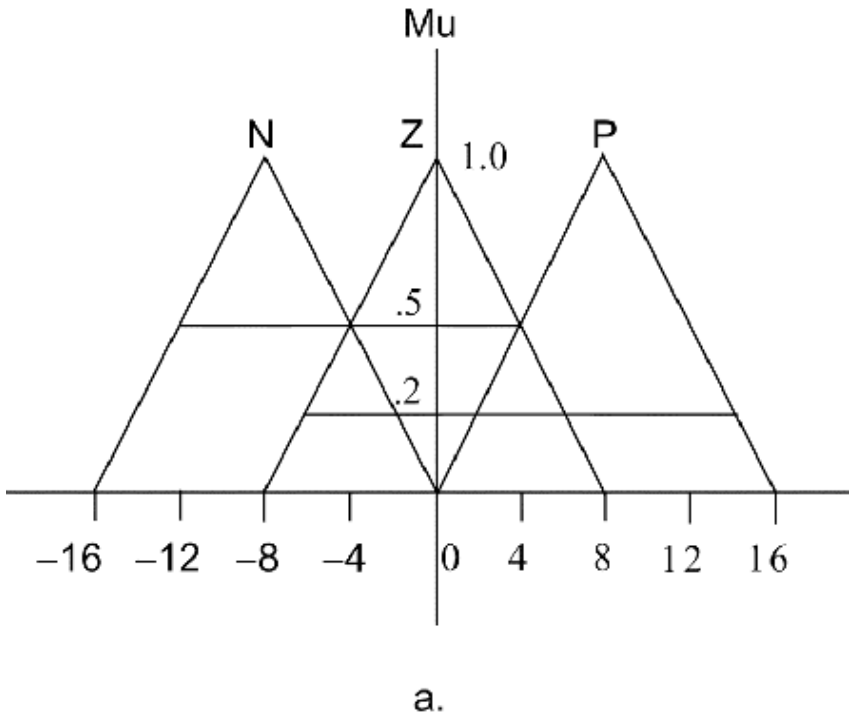


b.

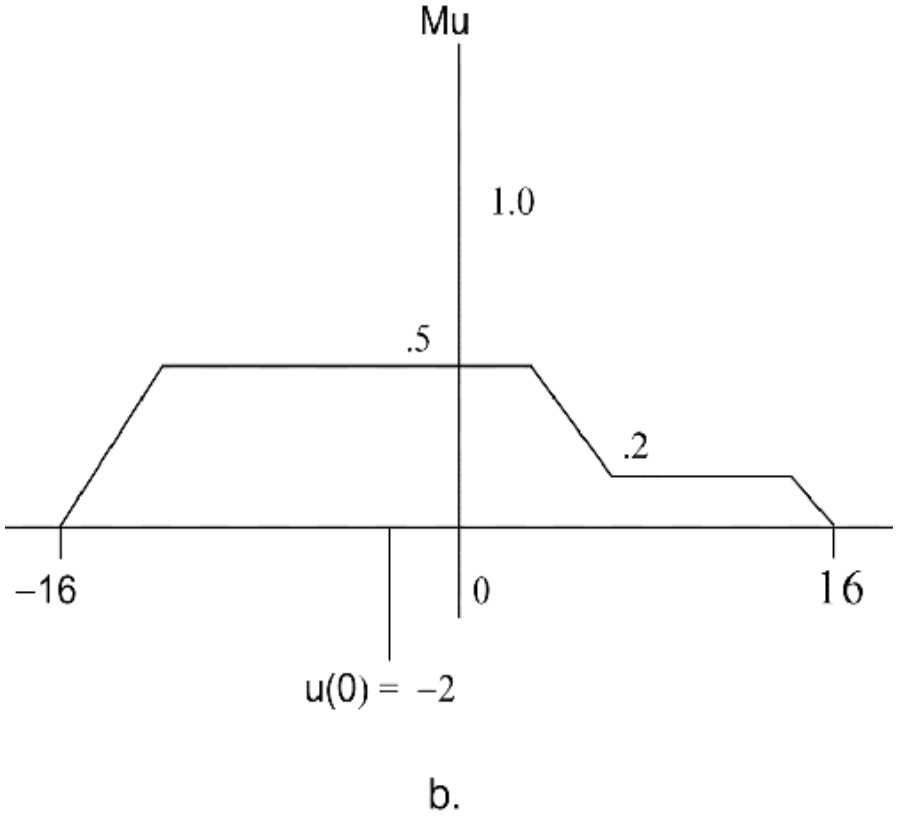
The Fuzzy Associative Matrix (FAM) for the pendulum problem

x_1 \ x_2	P	Z	N
P	PB	P	Z
Z	P	Z	N
N	Z	N	NB

The fuzzy consequents (a), and their union (b)



a.



b.

The centroid of the union (-2) is the crisp output.

Minimum of their measures is taken as the measure of the rule result

IF x_1 = P AND x_2 = Z THEN u = P

min(0.5, 0.2) = 0.2 P

IF x_1 = P AND x_2 = N THEN u = Z

min(0.5, 0.8) = 0.5 Z

IF x_1 = Z AND x_2 = Z THEN u = Z

min(0.5, 0.2) = 0.2 Z

IF x_1 = Z AND x_2 = N THEN u = N

min(0.5, 0.8) = 0.5 N

Procedure for control

- Take the crisp output and fuzzify it
- Check the Fuzzy Associative Matrix (FAM) to see which rules fire
  (4 rules fire in the example)
- Find the rule results
  - ANDed premises: take minimum
  - ORed premises: take maximum
- Combine the rule results
  (union in the example)
- Defuzzify to obtain the crisp output
  (centroid in the example)

Comments

- "fuzzy" refers to sets (as opposed to *crisp* sets)
- Fuzzy logic is useful in engineering control where the measurements are imprecise
- It has been successful in commercial control applications: automatic transmissions, trains, video cameras, electric shavers
- useful when there are small rule bases, no chaining of inferences, tunable parameters
- The theory is not concerned about how the rules are created, but how they are combined
- The rules are not chained together, instead all fire and the results are combined