Abduction, Uncertainty, and Probabilistic Reasoning

Chapter 15 and more

Introduction

- **Abduction** is a reasoning process that tries to form plausible explanations for abnormal observations
  - Abduction is distinct different from deduction and induction
  - Abduction is inherently uncertain
- Uncertainty becomes an important issue in AI research
- Some major formalisms for representing and reasoning about uncertainty
  - Mycin's certainty factor (an early representative)
  - Probability theory (esp. Bayesian belief networks)
  - Dempster-Shafer theory
  - Fuzzy logic
  - Truth maintenance systems

Abduction

- **Definition** (Encyclopedia Britannica): reasoning that derives an explanatory hypothesis from a given set of facts
  - The inference result is a *hypothesis*, which if true, could *explain* the occurrence of the given facts
- **Examples**
  - Dendral, an expert system to construct 3D structure of chemical compounds
    - Fact: mass spectrometer data of the compound and its chemical formula
    - KB: chemistry, esp. strength of different types of bounds
    - Reasoning: form a hypothetical 3D structure which meet the given chemical formula, and would most likely produce the given mass spectrum if subjected to electron beam bombardment

- Medical diagnosis
  - Facts: symptoms, lab test results, and other observed findings (called manifestations)
  - KB: causal associations between diseases and manifestations
  - Reasoning: one or more diseases whose presence would causally explain the occurrence of the given manifestations
- Many other reasoning processes (e.g., word sense disambiguation in natural language process, image understanding, detective's work, etc.) can also been seen as abductive reasoning.

Comparing abduction, deduction and induction

Deduction:	major premise:	All balls in the box are black	**A => B** **A** **---------** **B**
	minor premise:	These balls are from the box	
	conclusion:	These balls are black	
Abduction:	rule:	All balls in the box are black	**A => B** **B** **-------------** **Possibly A**
	observation:	These balls are black	
	explanation:	These balls are from the box	
Induction:	case:	These balls are from the box	**Whenever A then B but not vice versa** **------------** **Possibly A => B**
	observation:	These balls are black	
	hypothesized rule:	All ball in the box are black	

Induction: from specific cases to general rules

Abduction and deduction:

both from part of a specific case to other part of the case using general rules (in different ways)

Characteristics of abduction reasoning

1. Reasoning results are **hypotheses**, not theorems (may be false *even if* rules and facts are true),
   – e.g., misdiagnosis in medicine
2. There may be multiple plausible hypotheses
   – When given rules A => B and C => B, and fact B both A and C are plausible hypotheses
   – Abduction is inherently uncertain
   – Hypotheses can be ranked by their plausibility if that can be determined
3. Reasoning is often a Hypothesize- and-test cycle
   – hypothesize phase: postulate possible hypotheses, each of which could explain the given facts (or explain most of the important facts)
   – test phase: test the plausibility of all or some of these hypotheses

- One way to test a hypothesis H is to query if some thing that is currently unknown but can be predicted from H is actually true.
    - If we also know A => D and C => E, then ask if D and E are true.
    - If it turns out D is true and E is false, then hypothesis A becomes more plausible (support for A increased, support for C decreased)

4. Reasoning is non-monotonic
    - Plausibility of hypotheses can increase/decrease as new facts are collected (deductive inference determines if a sentence is true but would never change its truth value)
    - Some hypotheses may be discarded, and new ones may be formed when new observations are made

Source of Uncertainty

- Uncertain data (noise)
- Uncertain knowledge (e.g, causal relations)
  - A disorder may cause any and all POSSIBLE manifestations in a specific case
  - A manifestation can be caused by more than one POSSIBLE disorders
- Uncertain reasoning results
  - Abduction and induction are inherently uncertain
  - Default reasoning, even in deductive fashion, is uncertain
  - Incomplete deductive inference may be uncertain

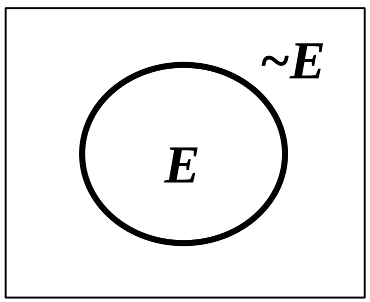
Probabilistic Inference

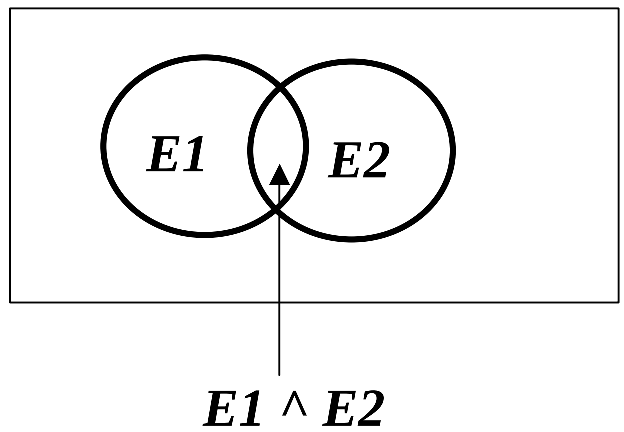
- Based on probability theory (especially Bayes' theorem)
  - Well established discipline about uncertain outcomes
  - Empirical science like physics/chemistry, can be verified by experiments
- Probability theory is too rigid to apply directly in many applications
  - Some assumptions have to be made to simplify the reality
  - Different formalisms have been developed in which some aspects of the probability theory are changed/modified.
- We will briefly review the basics of probability theory before discussing different approaches to uncertainty
- The presentation uses diagnostic process (an abductive and evidential reasoning process) as an example

Probability of Events

- Sample space and events
  - Sample space S: (e.g., all people in an area)
  - Events $E1 \subseteq S$: (e.g., all people having cough)
    $E2 \subseteq S$: (e.g., all people having cold)
- Prior (marginal) probabilities of events
  - $P(E) = |E| / |S|$ (frequency interpretation)
  - $P(E) = 0.1$ (subjective probability)
  - $0 <= P(E) <= 1$ for all events
  - Two special events: $\emptyset$ and S: $P(\emptyset) = 0$ and $P(S) = 1.0$
- Boolean operators between events (to form compound events)
  - Conjunctive (intersection): $E1 \wedge E2$ ($E1 \cap E2$)
  - Disjunctive (union): $E1 \vee E2$ ($E1 \cup E2$)
  - Negation (complement): $\sim E$ ($E^C = S - E$)

- Probabilities of compound events
  - P(~E) = 1 – P(E) because P(~E) + P(E) =1
  - P(E1 v E2) = P(E1) + P(E2) – P(E1 ^ E2)
  - But how to compute the *joint probability* P(E1 ^ E2)?





- Conditional probability (of E1, given E2)
  - How likely E1 occurs in the subspace of E2

$$\boldsymbol{P}(\boldsymbol{E}1 \mid \boldsymbol{E}2) = \frac{|\boldsymbol{E}1 \wedge \boldsymbol{E}2|}{|\boldsymbol{E}2|} = \frac{|\boldsymbol{E}1 \wedge \boldsymbol{E}2| / |\boldsymbol{S}|}{|\boldsymbol{E}2| / |\boldsymbol{S}|} = \frac{\boldsymbol{P}(\boldsymbol{E}1 \wedge \boldsymbol{E}2)}{\boldsymbol{P}(\boldsymbol{E}2)}$$

$$\boldsymbol{P}(\boldsymbol{E}1 \wedge \boldsymbol{E}2) = \boldsymbol{P}(\boldsymbol{E}1 \mid \boldsymbol{E}2)\boldsymbol{P}(\boldsymbol{E}2)$$

- Independence assumption
  - Two events E1 and E2 are said to be independent of each other if $\boldsymbol{P}(\boldsymbol{E}1 \mid \boldsymbol{E}2) = \boldsymbol{P}(\boldsymbol{E}1)$ (given E2 does not change the likelihood of E1)
  - It can simplify the computation

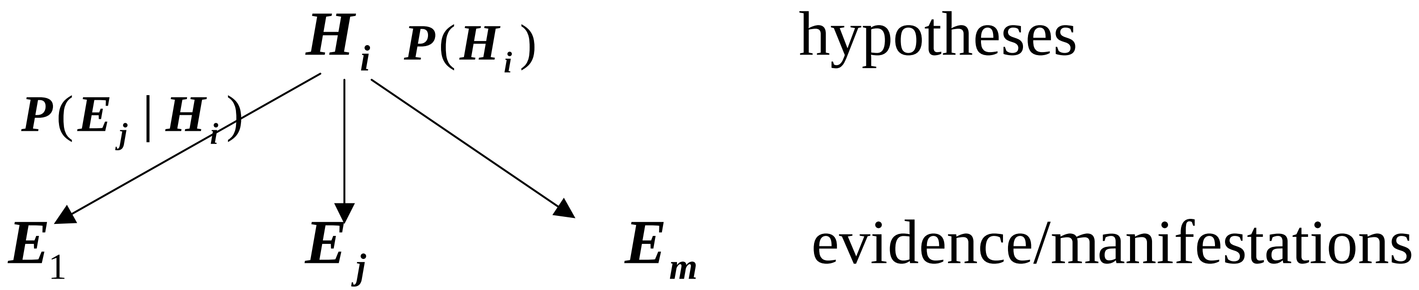
$$\boldsymbol{P}(\boldsymbol{E}1 \wedge \boldsymbol{E}2) = \boldsymbol{P}(\boldsymbol{E}1 \mid \boldsymbol{E}2)\boldsymbol{P}(\boldsymbol{E}2) = \boldsymbol{P}(\boldsymbol{E}1)\boldsymbol{P}(\boldsymbol{E}2)$$

$$\begin{aligned} \boldsymbol{P}(\boldsymbol{E}1 \vee \boldsymbol{E}2) &= \boldsymbol{P}(\boldsymbol{E}1) + \boldsymbol{P}(\boldsymbol{E}2) - \boldsymbol{P}(\boldsymbol{E}1 \wedge \boldsymbol{E}2) \\ &= \boldsymbol{P}(\boldsymbol{E}1) + \boldsymbol{P}(\boldsymbol{E}2) - \boldsymbol{P}(\boldsymbol{E}1)\boldsymbol{P}(\boldsymbol{E}2) \\ &= 1 - (1 - \boldsymbol{P}(\boldsymbol{E}1)(1 - \boldsymbol{P}(\boldsymbol{E}2)) \end{aligned}$$

- Mutually exclusive (ME) and exhaustive (EXH) set of events
  - ME: $\boldsymbol{E}_i \wedge \boldsymbol{E}_j = \varnothing \;\; (\boldsymbol{P}(\boldsymbol{E}_i \wedge \boldsymbol{E}_j) = 0), \boldsymbol{i}, \boldsymbol{j} = 1,..,\boldsymbol{n}, \boldsymbol{i} \neq \boldsymbol{j}$
  - EXH: $\boldsymbol{E}_1 \vee \ldots \vee \boldsymbol{E}_n = \boldsymbol{S} \;\; (\boldsymbol{P}(\boldsymbol{E}_1 \vee \ldots \vee \boldsymbol{E}_n) = 1)$

Bayes' Theorem

- In the setting of diagnostic/evidential reasoning



  – Know prior probability of hypothesis $P(H_i)$

  conditional probability $P(E_j \mid H_i)$

  – Want to compute the *posterior probability* $P(H_i \mid E_j)$

- Bayes' theorem (formula 1): $P(H_i \mid E_j) = P(H_i)P(E_j \mid H_i) / P(E_j)$
- If the purpose is to find which of the n hypotheses $H_1, \ldots, H_n$ is more plausible given E_j, then we can ignore the denominator and rank them use ***relative likelihood***

$$rel(H_i \mid E_j) = P(E_j \mid H_i)P(H_i)$$

- $P(E_j)$ can be computed from $P(E_j \mid H_i)$ and $P(H_i)$, if we assume all hypotheses $H_1, \ldots, H_n$ are ME and EXH

$$\begin{aligned} P(E_j) &= P(E_j \wedge (H_1 \vee \ldots \vee H_n) \qquad \text{(by EXH)} \\ &= \sum_{i=1}^{n} P(E_j \wedge H_i) \qquad \text{(by ME)} \\ &= \sum_{i=1}^{n} P(E_j \mid H_i) P(H_i) \end{aligned}$$

- Then we have another version of Bayes' theorem:

$$P(H_i \mid E_j) = \frac{P(E_j \mid H_i) P(H_i)}{\sum_{k=1}^{n} P(E_j \mid H_k) P(H_k)} = \frac{rel(H_i \mid E_j)}{\sum_{k=1}^{n} rel(H_k \mid E_j)}$$

where $\sum_{k=1}^{n} P(E_j \mid H_k) P(H_k)$, the sum of relative likelihood of all n hypotheses, is a normalization factor

Probabilistic Inference for simple diagnostic problems

- Knowledge base:

  $E_1,\ldots,E_m$: evidence/manifestation

  $H_1,\ldots,H_n$: hypotheses/disorders

  E_j and H_i are binary and hypotheses form a ME & EXH set

  $P(E_j \mid H_i), i=1,\ldots n, j=1,\ldots m$ conditional probabilities

- Case input: $E_1,\ldots,E_l$
- Find the hypothesis H_i with the highest posterior probability $P(H_i \mid E_1,\ldots,E_l)$
- By Bayes' theorem $P(H_i \mid E_1,\ldots,E_l) = \dfrac{P(E_1,\ldots E_l \mid H_i)P(H_i)}{P(E_1,\ldots E_l)}$
- Assume all pieces of evidence are conditionally independent, given any hypothesis

$$P(E_1,\ldots E_l \mid H_i) = \Pi_{j=1}^{l} P(E_j \mid H_i)$$

- The relative likelihood

$$rel(H_i \mid E_1,\ldots,E_l) = P(E_1,\ldots,E_l \mid H_i)P(H_i) = P(H_i)\prod_{j=1}^{l} P(E_j \mid H_i)$$

- The absolute posterior probability

$$P(H_i \mid E_1,\ldots,E_l) = \frac{rel(H_i \mid E_1,\ldots,E_l)}{\sum_{k=1}^{n} rel(H_k \mid E_1,\ldots,E_l)} = \frac{P(H_i)\prod_{j=1}^{l} P(E_j \mid H_i)}{\sum_{k=1}^{n} P(H_k)\prod_{j=1}^{l} P(E_j \mid H_k)}$$

- Evidence accumulation (when new evidence discovered)

$$rel(H_i \mid E_1,\ldots,E_l,E_{l+1}) = P(E_{l+1} \mid H_i)\,rel(H_i \mid E_1,\ldots,E_l)$$

$$rel(H_i \mid E_1,\ldots,E_l,\sim E_{l+1}) = (1 - P(E_{l+1} \mid H_i))\,rel(H_i \mid E_1,\ldots,E_l)$$

Assessment of Assumptions

- Assumption 1: hypotheses are mutually exclusive and exhaustive
  - Single fault assumption (one and only hypothesis must true)
  - **Multi-faults do exist in individual cases**
  - Can be viewed as an approximation of situations where hypotheses are independent of each other and their prior probabilities are very small

$$\boldsymbol{P}(\boldsymbol{H}_1 \wedge \boldsymbol{H}_2) = \boldsymbol{P}(\boldsymbol{H}_1)\boldsymbol{P}(\boldsymbol{H}_2) \approx 0 \text{ if both } \boldsymbol{P}(\boldsymbol{H}_1) \text{ and } \boldsymbol{P}(\boldsymbol{H}_2) \text{ are very small}$$

- Assumption 2: pieces of evidence are conditionally independent of each other, given any hypothesis
  - Manifestations themselves are not independent of each other, they are correlated by their common causes
  - Reasonable under single fault assumption
  - **Not so when multi-faults are to be considered**

Limitations of the simple Bayesian system

- Cannot handle well hypotheses of multiple disorders
  - Suppose $H_1, \ldots, H_n$ are independent of each other
  - Consider a composite hypothesis $H_1 \wedge H_2$
  - How to compute the posterior probability (or relative likelihood)

$$P(H_1 \wedge H_2 \mid E_1, \ldots, E_l)\ ?$$

  - Using Bayes' theorem

$$P(H_1 \wedge H_2 \mid E_1, \ldots, E_l) = \frac{P(E_1, \ldots E_l \mid H_1 \wedge H_2) P(H_1 \wedge H_2)}{P(E_1, \ldots E_l)}$$

$$P(H_1 \wedge H_2) = P(H_1) P(H_2) \text{ because they are independent}$$

$$P(E_1, \ldots E_l \mid H_1 \wedge H_2) = \Pi_{j=1}^{l} P(E_j \mid H_1 \wedge H_2)$$

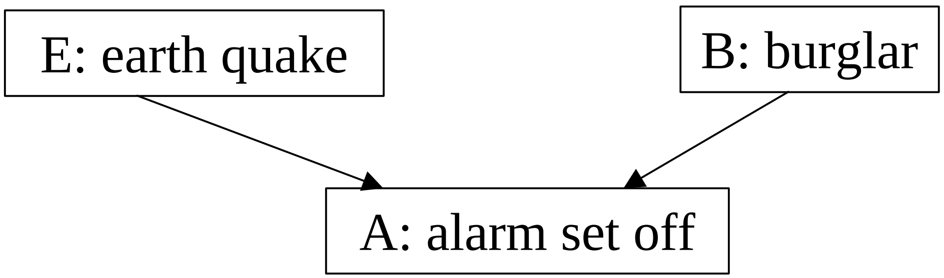
assuming E_j are independent, given $H_1 \wedge H_2$

How to compute $P(E_j \mid H_1 \wedge H_2)$?

– Assuming $\boldsymbol{H}_1, \ldots, \boldsymbol{H}_n$ are independent, given $\boldsymbol{E}_1, \ldots, \boldsymbol{E}_l$) ?

$$\boldsymbol{P}(\boldsymbol{H}_1 \wedge \boldsymbol{H}_2 \mid \boldsymbol{E}_1, \ldots, \boldsymbol{E}_l) = \boldsymbol{P}(\boldsymbol{H}_1 \mid \boldsymbol{E}_1, \ldots, \boldsymbol{E}_l)\, \boldsymbol{P}(\boldsymbol{H}_2 \mid \boldsymbol{E}_1, \ldots, \boldsymbol{E}_l)$$

but this is a very unreasonable assumption



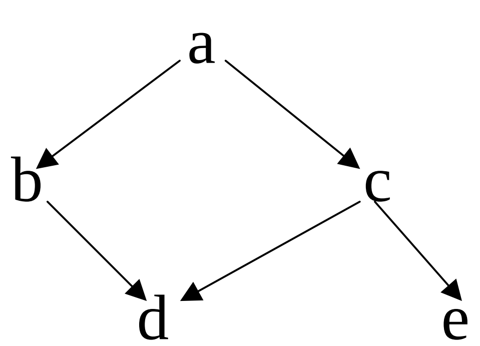
E and B are independent
But when A is given, they are (adversely) dependent because they become competitors to explain A

P(B|A, E) <<P(B|A)

- Cannot handle causal chaining
  - Ex. A: weather of the year
    B: cotton production of the year
    C: cotton price of next year
  - Observed: A influences C
  - The influence is not direct (A -> B -> C)
    P(C|B, A) = P(C|B): instantiation of B blocks influence of A on C
- Need a better representation and a better assumption

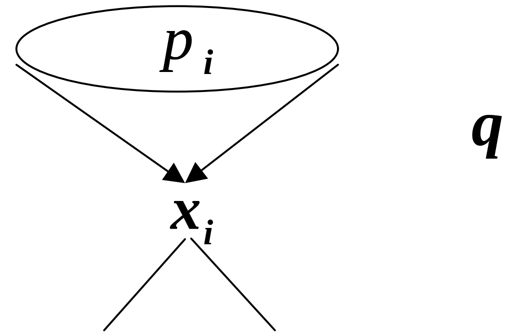
Bayesian Belief Networks (BBN)

- Definition: A BBN = (DAG, CPD)
  - **DAG**: directed acyclic graph
    nodes: random variables of interest (binary or multi-valued)
    arcs: direct causal/influential relations between nodes
  - **CPD**: conditional probability distribution at each node $\boldsymbol{x_i}$
    $\boldsymbol{P}(\boldsymbol{x_i} \mid p_i)$ where p_i is the set of all parent nodes of $\boldsymbol{x_i}$
  - For root nodes $p_i = \varnothing$, so $\boldsymbol{P}(\boldsymbol{x_i} \mid p_i) = \boldsymbol{P}(\boldsymbol{x_i})$
    Since roots are not influenced by anyone, they are considered independent of each other
- Example BBN



P(A) = 0.001
P(B|A) = 0.3 P(B|~A) = 0.001
P(C|A) = 0.2 P(C|~A) = 0.005
P(D|B,C) = 0.1 P(D|B,~C) = 0.01
P(D|~B,C) = 0.01 P(D|~B,~C) = 0.00001
P(E|C) = 0.4 P(E|~C) = 0.002

- Independence assumption
  - $\boldsymbol{P}(\boldsymbol{x_i} \mid p_{\boldsymbol{i}}, \boldsymbol{q}) = \boldsymbol{P}(\boldsymbol{x_i} \mid p_{\boldsymbol{i}})$
    where $\boldsymbol{q}$ is any set of variables (nodes) other than $\boldsymbol{x_i}$ and its successors
  - $p_{\boldsymbol{i}}$ blocks influence of other nodes on $\boldsymbol{x_i}$ and its successors (q influences $\boldsymbol{x_i}$ only through variables in $p_{\boldsymbol{i}}$)
  - With this assumption, the complete joint probability distribution of all variables in the network can be represented by (recovered from) local CPD by chaining these CPD



$$\boldsymbol{P}(\boldsymbol{x}_1, \ldots, \boldsymbol{x_n}) = \Pi_{\boldsymbol{i}=1}^{\boldsymbol{n}} \boldsymbol{P}(\boldsymbol{x_i} \mid p_{\boldsymbol{i}})$$

	P(A, B, C, D, E)	
=	P(E|A, B, ***C***, D) P(A, B, C, D)	by Bayes' theorem
=	P(E|C) P(A, B, C, D)	by indep. assumption
=	P(E|C) P(D|A, ***B, C***) P(A, B, C)	
=	P(E|C) P(D|B, C) P(C|***A,*** B) P(A, B)	
=	P(E|C) P(D|B, C) P(C|A) P(B|A) P(A)	

Inference with BBN

- **Belief update**
  - Original belief (no variable is instantiated) : the prior probability $\boldsymbol{P(x_i)}$
    If $\boldsymbol{x_i}$ is a root, then $\boldsymbol{P(x_i)}$ is given in BBN.
    Otherwise, $\boldsymbol{P(x_i)} = \sum_{p_i} \boldsymbol{P(x_i \mid p_i) P(p_i)}$

    $\boldsymbol{P(x_i \mid p_i)}$ is given, but computer $\boldsymbol{P(p_{\mathrm{i}})}$ is complicated
    Ex : $\boldsymbol{P(B,C) = P(A,B,C) + P(\sim A,B,C)}$

    $$\boldsymbol{= P(B \mid A,C)P(A,C) + P(B \mid \sim A,C)P(\sim A,C)}$$

    $$\boldsymbol{= P(B \mid A)P(C \mid A)P(A) + P(B \mid \sim A)P(C \mid \sim A)P(\sim A)}$$

  - When some variables are instantiated (say $\boldsymbol{x_j}$ has value $\boldsymbol{X_j}$),
    beliefs on all other variable $\boldsymbol{x_i}$ is changed to $\boldsymbol{P(x_i \mid X_j)}$

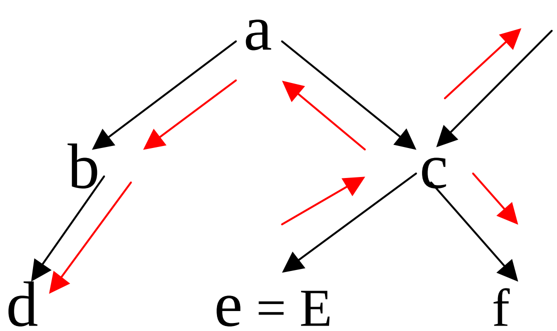
    $\boldsymbol{P(x_i \mid X_j)}$ can be computed from the joint probability distribution
    Ex : $\boldsymbol{d = D}$ and $\boldsymbol{e = E}$

    $$\boldsymbol{P(A \mid D,E) = \frac{P(A,D,E)}{P(D,E)} = \frac{\sum_{b,c} P(A,b,c,D,E)}{\sum_{a,b,c} P(a,b,c,D,E)}}$$

    This approach is not computationally feasible with large network

– ***Algorithmic approach*** (Pearl and others)

- Singly connected network, SCN (also known as poly tree)
  there is at most one undirected path between any two nodes
  (i.e., the network is a tree if the direction of arcs are ignored)
- The influence of the instantiated variable spreads to the rest of the network along the arcs
  – The instantiated variable influences its predecessors and successors differently
  – Computation is linear to the diameter of the network (the longest undirected path)



- For non-SCN (network with general structure)
  – Conditioning: find the the network's smallest cutset *C* (a set of nodes whose removal will render the network singly connected)
    for each instantiation of *C*, compute the belief update with the SCN algorithm
  – Combine the results from all possible instantiation of *C*.
  – Computationally expensive (finding the smallest cutset is itself NP-hard, and total number of possible instantiations of *C* is O(2^|*C*|.)

– ***Stochastic simulation***

- Randomly generate large number of instantiations of ALL variables $\boldsymbol{I}_k^{(n)}$ according to CPD
- Only keep those instantiations $\boldsymbol{I}_k^{(n)}$ which are consistent with the values of given instantiated variables
- Updated belief of those un-instantiated variables as their frequencies in the pool of recorded $\boldsymbol{I}_k^{(n)}$
- The accuracy of the results depend on the size of the pool (asymptotically approaches the exact results)

- **MAP problems**
  - Let $\boldsymbol{X}$ denote the set of all variables in a BBN, $\boldsymbol{V} \subseteq \boldsymbol{X}$ the set of instantiated variables, $\boldsymbol{U} = \boldsymbol{X} - \boldsymbol{V}$ the set of all un-instantiated varialbes. Then the MAP (*maximum aposteriori probability*) problem is to find the most probable instantiation of $\boldsymbol{U}$, given $\boldsymbol{V}$, i.e.,

    $$\max_{\boldsymbol{u}}(\boldsymbol{P}(\boldsymbol{U} \mid \boldsymbol{V}))$$

  - This is an optimization problem
  - Algorithms developed for ***exact*** solutions for different special BBN (Peng, Cooper, Pearl) have exponential complexity
  - Other techniques for ***approximate*** solutions
    - Genetic algorithms
    - Neural networks
    - Simulated annealing
    - Mean field theory

Noisy-Or BBN

- A special BBN of binary variables (Peng & Reggia, Cooper)
  - Each link $x_i \rightarrow x_j$ is associated with a probability value called ***causal strength*** c_{ij} that measures the strength of x_i ***alone*** may cause x_j, i.e., $c_{ij} = P(x_i \mid x_j$ is true and all others in p_i are false)
  - Causation independence: parent nodes influence a child independently
- Advantages:
  - One-to-one correspondence between causal links and causal strengths
  - Easy for humans to understand (acquire and evaluate KB)
  - Fewer # of probabilities needed in KB

$$\text{Complete joint prob. distribution}: \quad 2^{\mathrm{n}}$$

$$\text{General BBN}: \quad \sum_{i=1}^{n} 2^{|p_i|}$$

$$\text{Noisy-Or BBN}: \quad \sum_{i=1}^{n} |p_i|$$

  - Computation is less expensive
- Disadvantage: less expressive (less general)

Learning BBN (from case data)

- Need for learning
  - Experts' opinions are often biased, inaccurate, and incomplete
  - Large databases of cases become available
- What to learn
  - Learning CPD when DAG is known (easy)
  - Learning DAG (hard)
- Difficulties in learning DAG from case data
  - There are too many possible DAG when # of variables is large (more than exponential)

    n = 3, # of possible DAG = 25

    n = 10, # of possible DAG = 4*10^18
  - Missing values in database
  - Noisy data

- **Approaches**
  - ***Early effort***: Based on variable dependencies (Pearl)
    - Find all pairs of variables that are dependent of each other (applying standard statistical method on the database)
    - Eliminate (as much as possible) indirect dependencies
    - Determine directions of dependencies
    - Learning results are often incomplete (learned BBN contains indirect dependencies and undirected links)
  - ***Bayesian approach*** (Cooper)
    - Find the most probable DAG, given database DB, i.e.,
      max($\boldsymbol{P(DAG|DB)}$) or max($\boldsymbol{P(DAG, DB)}$)
    - Based on some assumptions, a formula is developed to compute ***P(DAG, DB)*** for a given pair of DAG and DB
    - A hill-climbing algorithm (K2) is developed to search a (sub)optimal DAG
    - Extensions to handle some form of missing values

– ***Minimum description length*** (MDL) (Lam)

- Sacrifices accuracy for simpler (less dense) structure
  – Case data not always accurate
  – Fewer links imply smaller CPD tables and less expensive inference
- ***L = L1 + L2*** where
  – ***L1***: the length of the encoding of DAG (smaller for simpler DAG)
  – ***L2***: the length of the encoding of the difference between DAG and DB (smaller for better match of DAG with DB)
  – Smaller ***L2*** implies more accurate (and more complex) DAG, and thus larger ***L1***
- Find DAG by heuristic best-first search, that Minimizes ***L***

– ***Neural network approach*** (Neal, Peng)

- For noisy-or BBN
- Maximizing $\boldsymbol{L} = \ln \prod_{\boldsymbol{V}^r \in \boldsymbol{D}} \boldsymbol{P}(\tilde{\boldsymbol{V}} = \boldsymbol{V}^r)$ where

  $\boldsymbol{D}$: case database; $\boldsymbol{V}^r$: case in $\boldsymbol{D}$; $\tilde{\boldsymbol{V}}$: state vector of the learned network

  $\boldsymbol{L}$ measures the similarity of the two distributions: one in $\boldsymbol{D}$, another in the learned network

Dempster-Shafer theory

- A variation of Bayes' theorem to represent **ignorance**
- Uncertainty and ignorance
  - Suppose two events A and B are ME and EXH, given an evidence E

    A: having cancer B: not having cancer E: smoking
  - By Bayes' theorem: our beliefs on A and B, given E, are measured by P(A|E) and P(B|E), and P(A|E) + P(B|E) = 1
  - In reality,

    I may have some belief in A, given E

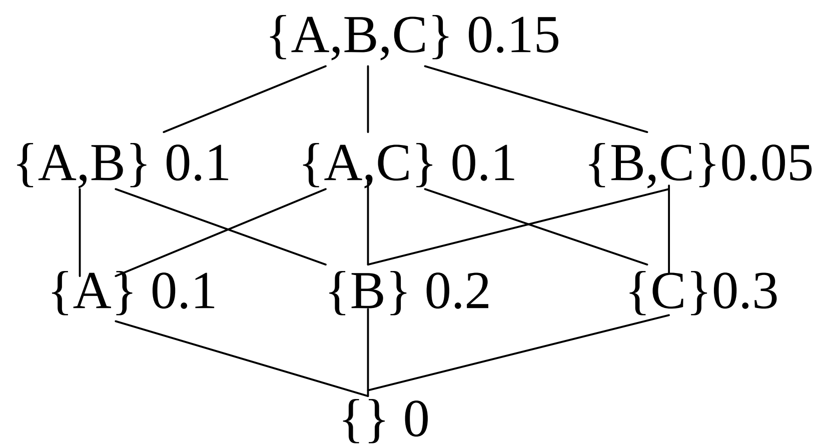
    I may have some belief in B, given E

    I may have some belief not committed to either one,
  - The uncommitted belief (ignorance) should not be given to either A or B, even though I know one of the two must be true, but rather it should be given to "A or B", denoted {A, B}
  - Uncommitted belief may be given to A and B when new evidence is discovered

- Representing ignorance
  - Frame of discernment : $q = \{\boldsymbol{h}_1, ..., \boldsymbol{h}_n\}$, a set of ME and EXH hypotheses. The power set 2^q is organized as a lattice of super/subset relations. Each node $\boldsymbol{S}$ is a subset of hypotheses ($\boldsymbol{S} \subseteq q$)
  - Ex: θ = {A,B,C}

  Each node $\boldsymbol{S}$ is associated with a basic probability assignment $\boldsymbol{m}(\boldsymbol{S})$

  $$0 \le \boldsymbol{m}(\boldsymbol{S}) \le 1;$$
  $$\boldsymbol{m}(\emptyset) = 0;$$
  $$\sum_{S \subseteq q} \boldsymbol{m(S)} = 1$$



- Belief function

$$\boldsymbol{Bel}(\boldsymbol{S}) = \sum_{\boldsymbol{S}' \subseteq \boldsymbol{S}} \boldsymbol{m}(\boldsymbol{S}'); \quad \boldsymbol{Bel}(\emptyset) = 0; \quad \boldsymbol{Bel}(q) = 1$$
$$\boldsymbol{Bel}(\{\boldsymbol{A}, \boldsymbol{B}\}) = \boldsymbol{m}(\{\boldsymbol{A}, \boldsymbol{B}\}) + \boldsymbol{m}(\{\boldsymbol{A}\}) + \boldsymbol{m}(\{\boldsymbol{B}\}) + \boldsymbol{m}(\emptyset)$$
$$= 0.1 + 0.1 + 0.2 + 0 = 0.4$$
$$\boldsymbol{Bel}(\{\boldsymbol{A}, \boldsymbol{B}\}^{\boldsymbol{C}}) = \boldsymbol{Bel}(\{\boldsymbol{C}\}) = 0.3$$

– Plausibility (upper bound of belief of a node)

All belief not committed to $\boldsymbol{S}^{\boldsymbol{C}}$ may be commited to $\boldsymbol{S}$

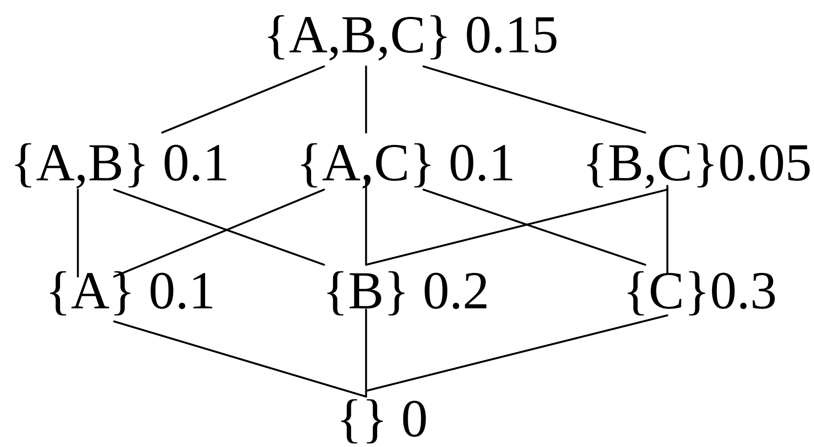
$$\boldsymbol{Pls}(\boldsymbol{S}) = 1 - \boldsymbol{Bel}(\boldsymbol{S}^{\boldsymbol{C}})$$

$$\boldsymbol{Pls}(\{\boldsymbol{A},\boldsymbol{B}\}) = 1 - \boldsymbol{Bel}(\{\boldsymbol{C}\}) = 1 - 0.3 = 0.7$$

$$[\boldsymbol{Bel}(\boldsymbol{S}), \quad \boldsymbol{Pls}(\boldsymbol{S})] \text{ belief interval}$$

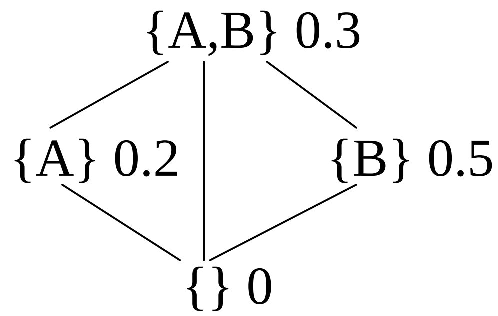
Lower bound (known belief)

Upper bound (maximally possible)



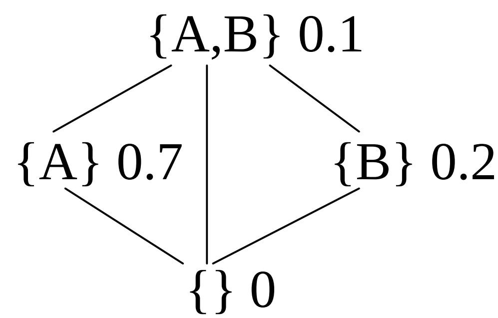
- **Evidence combination** (how to use D-S theory)
  - Each piece of evidence has its own m(.) function for the same θ

    $q = \{\boldsymbol{A}, \boldsymbol{B}\}$: A : having cancer; B : not having cancer



$\boldsymbol{m}_1(\boldsymbol{S})$

$\boldsymbol{E}_1$: smoking



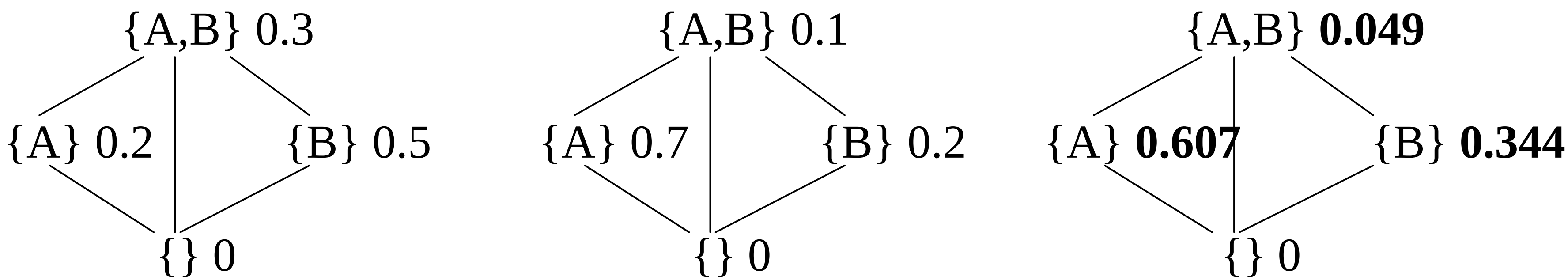
$\boldsymbol{m}_2(\boldsymbol{S})$

$\boldsymbol{E}_2$: living in high radiation area

  - Belief based on combined evidence can be computed from

$$\boldsymbol{m}(\boldsymbol{S}) = \boldsymbol{m}_1(\boldsymbol{S}) \oplus \boldsymbol{m}_2(\boldsymbol{S}) = \frac{\sum_{\boldsymbol{X} \cap \boldsymbol{Y} = \boldsymbol{S}} \boldsymbol{m}_1(\boldsymbol{X})\boldsymbol{m}_2(\boldsymbol{Y})}{1 - \underbrace{\sum_{\boldsymbol{X} \cap \boldsymbol{Y} = \varnothing} \boldsymbol{m}_1(\boldsymbol{X})\boldsymbol{m}_2(\boldsymbol{Y})}_{\text{incompatible combination}}}$$

normalization factor



E1 E2 E1 ^ E2

$$\boldsymbol{m}(\{\boldsymbol{A}\}) = \frac{\boldsymbol{m}_1(\{\boldsymbol{A}\})\boldsymbol{m}_2(\{\boldsymbol{A}\}) + \boldsymbol{m}_1(\{\boldsymbol{A}\})\boldsymbol{m}_2(\{\boldsymbol{A},\boldsymbol{B}\}) + \boldsymbol{m}_1(\{\boldsymbol{A},\boldsymbol{B}\})\boldsymbol{m}_2(\{\boldsymbol{A}\})}{1-[\boldsymbol{m}_1(\{\boldsymbol{A}\})\boldsymbol{m}_2(\{\boldsymbol{B}\}) + \boldsymbol{m}_1(\{\boldsymbol{B}\})\boldsymbol{m}_2(\{\boldsymbol{A}\})]}$$

$$= \frac{0.2\cdot 0.7 + 0.2\cdot 0.1 + 0.3\cdot 0.7}{1-[0.2\cdot 0.2 + 0.5\cdot 0.7]} = \frac{0.37}{0.61} = 0.607$$

$$\boldsymbol{m}(\{\boldsymbol{B}\}) = \frac{\boldsymbol{m}_1(\{\boldsymbol{B}\})\boldsymbol{m}_2(\{\boldsymbol{B}\}) + \boldsymbol{m}_1(\{\boldsymbol{B}\})\boldsymbol{m}_2(\{\boldsymbol{A},\boldsymbol{B}\}) + \boldsymbol{m}_1(\{\boldsymbol{A},\boldsymbol{B}\})\boldsymbol{m}_2(\{\boldsymbol{B}\})}{1-[\boldsymbol{m}_1(\{\boldsymbol{A}\})\boldsymbol{m}_2(\{\boldsymbol{B}\}) + \boldsymbol{m}_1(\{\boldsymbol{B}\})\boldsymbol{m}_2(\{\boldsymbol{A}\})]}$$

$$= \frac{0.5\cdot 0.2 + 0.5\cdot 0.1 + 0.3\cdot 0.2}{1-[0.2\cdot 0.2 + 0.5\cdot 0.7]} = \frac{0.21}{0.61} = 0.344$$

$$\boldsymbol{m}(\{\boldsymbol{A},\boldsymbol{B}\}) = \frac{\boldsymbol{m}_1(\{\boldsymbol{A},\boldsymbol{B}\})\boldsymbol{m}_2(\{\boldsymbol{A},\boldsymbol{B}\})}{0.61} = \frac{0.03}{0.61} = 0.049$$

– Ignorance is reduced

from m1({A,B}) = 0.3 to m({A,B}) = 0.049)

– Belief interval is narrowed

A: from [0.2, 0.5] to [0.607, 0.656]

B: from [0.5, 0.8] to [0.344, 0.393]

- Advantage:
  - The only formal theory about ignorance
  - Disciplined way to handle evidence combination
- Disadvantages
  - Computationally very expensive (lattice size $2^{|\theta|}$)
  - Assuming hypotheses are ME and EXH
  - How to obtain m(.) for each piece of evidence is not clear, except subjectively

Fuzzy sets and fuzzy logic

- Ordinary set theory
  - $$f_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases}$$

    $f_A(x)$ is called the characteristic or membership function of set A

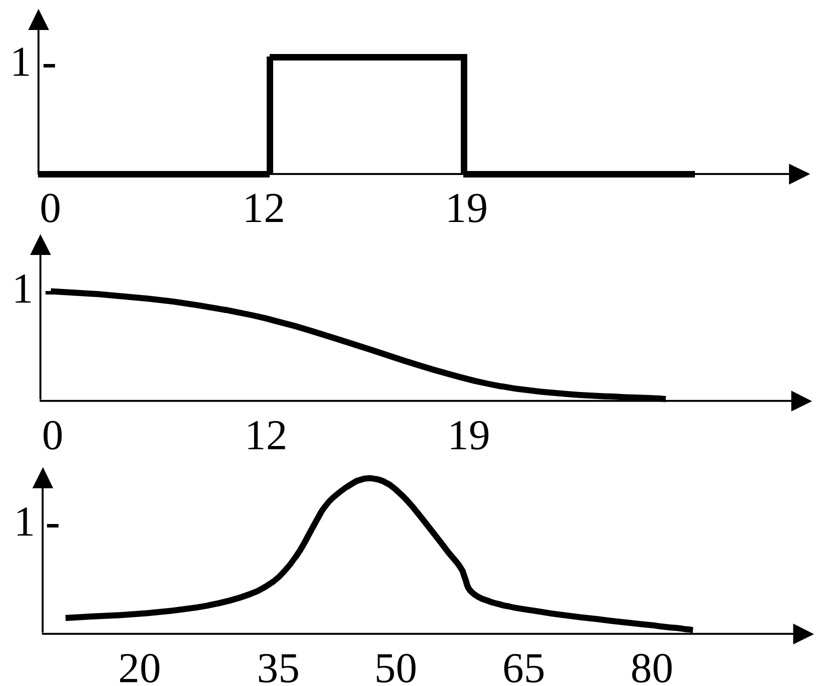
    $$\text{Predicate } A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases}$$

    When it is uncertain if $x \in A$, use probability $P(x \in A)$
  - There are sets that are described by vague linguistic terms (sets without hard, clearly defined boundaries), e.g., tall-person, fast-car
    - Continuous
    - Subjective (context dependent)
    - Hard to define a clear-cut 0/1 membership function

- Fuzzy set theory
  - Relax $f_A(x)$ from binary $\{0,1\}$ to continuous $[0,1]$ stands for the degree x is thought to belong to set A

    height(john) = 6’5”   Tall(john) = 0.9
    height(harry) = 5’8”   Tall(harry) = 0.5
    height(joe) = 5’1”   Tall(joe) = 0.1

  - Examples of membership functions



Set of teenagers

Set of young people

Set of mid-age people

- Fuzzy logic: many-value logic
  - Fuzzy predicates (degree of truth) $\boldsymbol{F_A}(\boldsymbol{x}) = \mathrm{y}$ if $\boldsymbol{f_A}(\boldsymbol{x}) = \mathrm{y}$
  - Connectors/Operators

$$\text{negation}: \boldsymbol{\sim F_A}(\boldsymbol{x}) = 1 - \boldsymbol{F_A}(\boldsymbol{x})$$
$$\text{conjunction}: \boldsymbol{F_A}(\boldsymbol{x}) \wedge \boldsymbol{F_B}(\boldsymbol{x}) = \min\{\boldsymbol{F_A}(\boldsymbol{x}), \boldsymbol{F_B}(\boldsymbol{x})\}$$
$$\text{disjunction}: \boldsymbol{F_A}(\boldsymbol{x}) \vee \boldsymbol{F_B}(\boldsymbol{x}) = \max\{\boldsymbol{F_A}(\boldsymbol{x}), \boldsymbol{F_B}(\boldsymbol{x})\}$$

- Compare with probability theory
  - Prob. Uncertainty of outcome,
    - Based on large # of repetitions or instances
    - For each experiment (instance), the outcome is either true or false (without uncertainty or ambiguity)
      unsure before it happens but sure after it happens

    Fuzzy: vagueness of conceptual/linguistic characteristics
    - Unsure even after it happens
      whether a child of tall mother and short father is tall
      unsure before the child is born
      unsure after grown up (height = 5’6”)

- Empirical vs subjective (testable vs agreeable)
- Fuzzy set connectors may lead to unreasonable results
  - Consider two events A and B with $P(A) < P(B)$
  - If $A \Rightarrow B$ (or $A \subseteq B$) then

    $P(A \wedge B) = P(A) = \min\{P(A), P(B)\}$

    $P(A \vee B) = P(B) = \max\{P(A), P(B)\}$
  - Not the case in general

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

    $P(A \vee B) = P(A) + P(B) - P(A \wedge B) \geq P(B)$

    (equality holds only if $P(B|A) = 1$, i.e., $A \Rightarrow B$)
- Something prob. theory cannot represent
  - Tall(john) = 0.9, ~Tall(john) = 0.1

    Tall(john) ^ ~Tall(john) = min{0.1, 0.9) = 0.1

    john's degree of membership in the fuzzy set of "median-height people" (both Tall and not-Tall)
  - In prob. theory: $P(john \in Tall \wedge john \notin Tall) = 0$

Uncertainty in rule-based systems

- Elements in Working Memory (WM) may be uncertain because
  - Case input (initial elements in WM) may be uncertain
    Ex: the CD-Drive does not work 70% of the time
  - Decision from a rule application may be uncertain even if the rule's conditions are met by WM with certainty
    Ex: flu => sore throat with high probability
- Combining symbolic rules with numeric uncertainty: Mycin's **Uncertainty Factor (CF)**
  - An early attempt to incorporate uncertainty into KB systems
  - CF $\in$ [-1, 1]
  - Each element in WM is associated with a CF: certainty of that assertion
  - Each rule *C1,...,Cn => Conclusion* is associated with a CF: certainty of the association (between *C1,...Cn* and *Conclusion*).

- CF propagation:
  - Within a rule: each *Ci* has CFi, then the certainty of Action is

    min{CF1,...CFn} * CF-of-the-rule
  - When more than one rules can apply to the current WM for the same *Conclusion* with different CFs, the ***largest of these CFs*** will be assigned as the CF for *Conclusion*
  - Similar to fuzzy rule for conjunctions and disjunctions
- Good things of Mycin's CF method
  - Easy to use
  - CF operations are reasonable in many applications
  - Probably the only method for uncertainty used in real-world rule-base systems
- Limitations
  - It is in essence an ad hoc method (it can be viewed as a probabilistic inference system with some strong, sometimes unreasonable assumptions)
  - May produce counter-intuitive results.