Uncertainty Reasoning

- "Nothing is certain but death and taxes"
  Benjamin Franklin

Uncertainty Reasoning

- Uncertainty
- Probability
- Syntax and Semantics
- Inference
- Independence and Bayes' Rule

Uncertainty

- So far our problems have assumed:
  - Start state is known with certainty
  - Actions are deterministic
  - Both assumptions are unrealistic (e.g. robotics)
- Knowledge:
  - Can a coffee delivery robot know a priori if coffee is made? Mail waits?
- Actions:
  - Robot grabs coffee: could fail (try again); could spill (make more)
  - Robot may move to office and may end up in lab.

Limitation of Deterministic Logic

- Pure logic fails for three main reasons:
- **Laziness:**
  - Too much work to list complete set of antecedents or consequents needed to ensure exception less rules, too hard to use the enormous rules that result
- **Theoretical ignorance:**
  - Science has no complete theory for the domain
- **Practical ignorance:**
  - Even if we know all the rules, we may be uncertain about a particular occurrence because all the necessary tests have not or cannot be run

Uncertainty

- Given action A_t – leave for airport t minutes before flight
- Will I catch my flight?

- Problems
  - Partial observability (road state, accidents, etc.)
  - Noisy sensors (radio traffic reports, smoke ahead, etc.)
  - Uncertainty in action outcomes (flat tire, etc.)
  - Immense complexity of modeling and predicting traffic

- A FOPC approach either:
  - Risks falsehood, or
  - Leads to conclusions that are too weak for decision making

Non-monotonic Logic

- Traditional logic is monotonic .
  - The set of legal conclusions grows monotonically with the set of facts appearing in our initial database.
- When humans reason, we use a defeasible logic.
  Almost every conclusion we draw is subject to reversal.
  If we find contradicting information later, we retract earlier inferences.

- Nonmonotonic logic, or Defeasible reasoning, allows a statement to be retracted.
- Solution: *Truth Maintenance*
  - Keep explicit information about which facts/inferences support other inferences.
  - If the foundation disappears, so must the conclusion.

Uncertainty

- On the other hand, the problem might not be the fact that T/F values can change over time, but rather that we are not certain of the T/F value.
- Agents almost never have access to the whole truth about their environment
- Agents must therefore act in the presence of uncertainty
  - Some information ascertained from facts
  - Some information inferred from facts and knowledge about environment
  - Some information is based on assumptions made from experience

Degrees of Belief and Preference

- The right decision requires a consideration of how important various objectives are, how likely they are to be achieved, and make tradeoffs between them.
- This generally requires that we quantify our preferences.
- We'll quantify our beliefs using probabilities
  - Pr(q) denotes probability that you believe q is true
  - Pr(A_{25}|no reported accidents) = 0.06
    denotes that given no reported accidents the probability of arriving on time if I leave 25 minutes before departure is 6%

Probability

- Probabilities relate propositions to one's own state of knowledge
- Probabilities are real values between 0.0 and 1.0 (inclusive) that represent ideal certainties of statements, given assumptions about the circumstances in which the statements apply.
- These values can be verified by testing, unlike certainty values. They apply in highly controlled situations.

$$Probability(event) = \Pr(event) = \frac{\#\text{ instances of the event}}{\text{total }\#\text{ of instances}}$$

Where do Probabilities Come From?

  - Frequency – primary method (Maximum Likelihood)
  - Subjective judgment
- Consider the probability that the sun will still exist tomorrow. There are several ways to compute this.
- Choice of experiment is know as the reference class problem

Example

- For example, if we roll two dice, each showing one of six possible numbers, the number of total unique rolls is 6*6 = 36. We distinguish the dice in some way (a first and second or left and right die). Here is a listing of the joint possibilities for the dice:
  - (1,1) (1,2) (1,3) (1,4) (1,5) (1,6) (2,1) (2,2) (2,3) (2,4) (2,5) (2,6) (3,1) (3,2) (3,3) (3,4) (3,5) (3,6) (4,1) (4,2) (4,3) (4,4) (4,5) (4,6) (5,1) (5,2) (5,3) (5,4) (5,5) (5,6) (6,1) (6,2) (6,3) (6,4) (6,5) (6,6)

- The number of rolls which add up to 4 is 3 ((1,3), (2,2), (3,1)), so the probability of rolling a total of 4 is 3/36 = 1/12.
  - This does not mean 8.3% true, but an 8.3% chance of it being true, or a probability of 0.083.

Probability Explanation

- Pr(*event*) is the probability in the absence of any additional information
- Probability depends on evidence
  - Before looking at dice: Pr(sum of 4) = 1/12
  - After looking at dice: Pr(sum of 4) = 0 or 1, depending on what we see

- All probability statements must indicate the evidence with respect to which the probability is being assessed.
  Pr(sum of 4| *evidence*)

- As new evidence is collected, probability calculates are updated.

- Before specific evidence is obtained, we refer to the prior or unconditional probability of the event with respect to the evidence. After the evidence is obtained, we refer to the posterior or conditional probability.

Making Decision Under Uncertainty

- Suppose I believe the following:
  $Pr(A_{25}| ..) = 0.04$
  $Pr(A_{90}| ..) = 0.70$
  $Pr(A_{120}| ..) = 0.95$
  $Pr(A_{1440}| ..) = 0.99$

- Which action to choose?
  - Probability theory tells us which is most likely
  - But depends on my preferences for missing flights vs. airport food, etc.
    - Utility theory is used to represent and infer preferences
    - Decision theory = utility theory + probability theory

Probability Basics

- From the set Ω - the sample space
  e.g., 6 possible rolls of a die
  $\omega \in \Omega$ is a sample point/ possible world/ atomic event

- A probability space or probability model is a sample space with an assignment $Pr(\omega)$ for every $\omega \in \Omega$ s.t.
  $0.0 \leq Pr(\omega) \leq 1.0$
  $\Sigma_\omega Pr(\omega) = 1.0$
  e.g., Pr(1) = Pr(2) = Pr(3) = Pr(4) = Pr(5) = Pr(6) = 1/6

- An event A is any set where $A \subseteq \Omega$
  $Pr(A) = \Sigma_{\{\omega \in A\}} Pr(\omega)$
  e.g., Pr(die roll < 4) = 1/6 + 1/6 + 1/6 = 1/2

Possible Worlds

- Think of a proposition as the event (set of sample points) where the proposition is true
- Given Boolean random variables *A* and *B*:
  event *a* = set of sample points where $A(\omega)$ = *true*
  event $\neg a$ = set of sample points where $A(\omega)$ = *false*
  event $a \wedge b$ = points where $A(\omega)$ = *true* and $B(\omega)$ = *true*
- Often in AI applications, the sample points are defined by the values of a set of random variables
- A formula is a logical combination of variable assignments:
  - X = x1; (X = x2 $\vee$ X = x3) $\wedge$ Y = y2
- A possible world is an assignment of values to each random variable.
  - These are analogous to truth assignments (interpretations)

Probability Distributions

- Prior or unconditional probabilities of propositions
  Pr(Cavity=*true*) = 0.1 and Pr(Weather=*sunny*) = 0.72
- If we want to know the probability of a variable that can take on multiple values, we define a probability distribution, or a set of probabilities for each possible variable value.
  Weather = (*sunny*, *rain*, *cloudy*, *snow*)
  Pr(Weather) = $\langle 0.7, 0.2, 0.08, 0.02 \rangle$
- Note that the sum of the probabilities for possible values of a variable must always sum to 1, and that Pr(α) is the sum of those worlds in which α is true.

$$\Pr(\alpha) = \sum_{\omega \in \Omega} \{\Pr(\omega) : \omega \models \alpha\}$$

Joint Probability Distributions

- Because events are rarely isolated from other events, we define a joint probability distribution which for a set of random variables gives the probability of every atomic event on those random variables
- The joint probability distribution is an *n*-dimensional array of combinations of probabilities for that state occurring

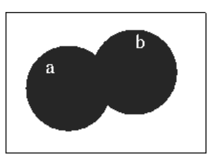
Weather =	*sunny*	*rain*	*cloudy*	*snow*
Cavity = *true*	0.144	0.02	0.016	0.02
Cavity = *false*	0.576	0.08	0.064	0.008

- Sum for a variable is unconditional Probability

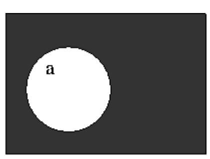
Every question about a domain can be answered by the joint distribution because every event is a sum of sample points

Axioms of Probability

- $0.0 \leq \Pr(event) \leq 1.0$
- Disjunction, $a \vee b$: $\Pr(a \vee b) = \Pr(a) + \Pr(b) - \Pr(a \wedge b)$

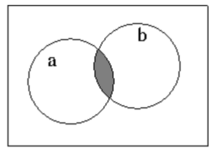


- Negation, $\Pr(\neg a) = 1 - \Pr(a)$



Axioms of Probability

- Conjunction Product Rule:
  - Pr($a \wedge b$) = Pr(b|a) * Pr(a)
    Pr($a \wedge b$) = Pr(a|b) * Pr(b)



- The only way *a* and *b* can both be true is if *a* is true and we know *b* is true given *a* is true (thus *b* is also true).

- If *a* and *b* are independent events (the truth of *a* has no effect on the truth of *b*, then Pr($a \wedge b$) = Pr(a) * Pr(b).
  - "Wet" and "Raining" are not independent events.
    "Wet" and "Joe made a joke" are pretty close to independent events.

Chain Rule

- The chain rule is derived by successive application of the product rule:

$$\Pr(X_1,\ldots,X_n) = \Pr(X_1,\ldots,X_{n-1})\Pr(X_n \mid X_1,\ldots,X_{n-1})$$
$$= \Pr((X_1,\ldots,X_{n-2})\Pr(X_{n-1} \mid X_1,\ldots,X_{n-2})\Pr(X_n \mid X_1,\ldots,X_{n-1})$$
$$= \ldots$$
$$= \prod_{i=1}^{n} \Pr(X_i \mid X_1,\ldots,X_{i-1})$$

- Summing Out Rule

$$\Pr(a) = \sum_{b \in Dom(B)} \Pr(a \mid b)\Pr(b)$$

Conditional Probability

- Once evidence is obtained, the agent can use conditional probabilities, Pr(*a*|*b*)
  - Pr(*a*|*b*) = probability of *a* being true given that we known *b* is true
  - The equation $\Pr(a \mid b) = \frac{\Pr(a \wedge b)}{\Pr(b)}$ holds whenever Pr(*b*) > 0

- An agent who bets according to probabilities that violate these axioms can be forced to bet so as to lose money regardless of outcome [deFinetti, 1931]

Axioms of Probability

- Bayes' Rule Given a hypothesis (H) and evidence (E), and given that P(E) $\neq$ 0, what is P(H|E)?

- Many times rules and information are uncertain, yet we still want to say something about the consequent; namely, the degree to which it can be believed. A British cleric and mathematician, Thomas Bayes, suggested an approach.

- Recall the two forms of the product rule:
  - $P(a \wedge b) = P(a) * P(b|a)$
    $P(a \wedge b) = P(b) * P(a|b)$

- If we equate the two right-hand sides and divide by P(a), we get Bayes' Rule:

$$P(b \mid a) = \frac{P(a \mid b) P(b)}{P(a)}$$

Example

- I have three identical boxes labeled H1, H2, and H3
  I place 1 black bead and 3 white beads into H1
  I place 2 black beads and 2 white beads into H2
  I place 4 black beads and no white beads into H3
  I draw a box at random, and remove a bead from that box. Given the color of the bead, which box did am I holding?
- If I replace the bead, then redraw another bead at random from the same box, how well can I predict its color before drawing it?
- These two questions are the foundation of uncertainty reasoning and machine learning.

Answer

- Observation: I draw a white bead.
  - $P(H_1|W) = P(H_1)P(W|H_1) / P(W)$
    $= (1/3 * 3/4) / 5/12 = 3/12 * 12/5 = 36/60 = 3/5$
  - $P(H_2|W) = P(H_2)P(W|H_2) / P(W)$
    $= (1/3 * 1/2) / 5/12 = 1/6 * 12/5 = 12/30 = 2/5$
  - $P(H_3|W) = P(H_3)P(W|H_3) / P(W)$
    $= (1/3 * 0) / 5/12 = 0 * 12/5 = 0$

Example

- Boxes H1, H2, and H3 were my prior models of the world
- The fact that P(H1) = 1/3, P(H2) = 1/3, and P(H3) = 1/3 (uniformly distributed) was my prior distribution
- The color of the bead was a piece of evidence about the true model of the world
- The use of Bayes' rule was a piece of probabilistic inference, giving me a posterior distribution on possible worlds
- Learning is prior + evidence $\rightarrow$ posterior
  Maximum A Posteriori (MAP) hypothesis
- A piece of evidence decreases my ignorance about the world
- Distributions are good ways of describing your state of knowledge. Knowledge that includes an uncertainty measure can mean much better decision making.

Example

- Bayes' rule is useful when we have three of the four parts of the equation. In this example, a doctor knows that meningitis causes a stiff neck in 50% of such cases. The prior probability of having meningitis is 1/50,000 and the prior probability of any patient having a stiff neck is 1/20. What is the probability that a patient has meningitis if they have a stiff neck?

H = "Patient has meningitis“   E = "Patient has stiff neck"

$$P(H|E) = \frac{P(E|H)*P(H)}{P(E)}$$

P(H|E) = (0.5*.00002)/.05 = .000

Inference by Enumeration

- Starting with the joint distribution:

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

- For any proposition ϕ, sum the atomic events where it is true: $\Pr(\phi) = \Sigma_{\omega|\omega\models\phi}\Pr(\omega)$
- $\Pr(\mathit{toothache})$= 0.108+0.012+0.016+0.064 = 0.2
- $\Pr(\mathit{cavity} \vee \mathit{toothache})$ = 0.2+0.072+0.008 = 0.28
- $\Pr(\neg\mathit{cavity}|\mathit{toothache}) = \dfrac{\Pr(\neg\mathit{cavity}\wedge\mathit{toothache})}{\Pr(\mathit{toothache})}$
  $= \dfrac{0.016+0.64}{0.108+0.012+0.016+0.064} = 0.4$

Normalization

- Starting with the joint distribution:

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

- $\Pr(\neg\mathit{cavity}|\mathit{toothache})$ = 0.4
- Denominator can be viewed as a normalizing constant α or η
- $\Pr(\mathit{cavity}|\mathit{toothache}) = \alpha\Pr(\mathit{cavity},\mathit{toothache})$
  $= \alpha\langle[\Pr(\mathit{cavity},\mathit{toothache},\mathit{catch})+\Pr(\mathit{cavity},\mathit{toothache},\neg\mathit{catch})],$
  $[\Pr(\neg\mathit{cavity},\mathit{toothache},\mathit{catch})+\Pr(\neg\mathit{cavity},\mathit{toothache},\neg\mathit{catch})]\rangle$
  $= \alpha\langle[0.108+0.012],[0.016+0.064]\rangle$
  $=\alpha\langle 0.12,0.08\rangle=\langle 0.60,0.40\rangle$

Lunar Lander Example

- A lunar lander crashes somewhere in your town (one of the cells at random in the grid). The crash point is uniformly random (the probability is uniformly distributed, meaning each location has an equal probability of being the crash point).

					D	D	D	
R	R	R	R	R	DR	DR	DR	R
R	R	R	R	R	DR	DR	DR	R
					D	D	D	

- D is the event that it crashes downtown.
- R is the event that it crashes in the river.

- What is P(R)? 18/54 = 0.333
- What is P(D)? 12/54 = 0.222
- What is P(D∧R)? 6/54 = 0.111
- What is P(D|R)?
- What is P(R|D)?
- What is P(R∧D)/P(D)?

Inference: Computational Bottleneck

- Issue1: how do we specify the full joint distribution over X_1, X_2,…,X_n?
  - Exponential number of possible worlds
    - e.g. if the X_i are boolean, then 2^n numbers
  - These numbers are not robust/stable
  - These numbers are not natural to assess (what is probability that there is a fire at home; it's raining in Tibet; robot charge level is low;...”?)

Inference: Computational Bottleneck

- Issue 2: Inference by enumeration is slow
  - Must sum over exponential number of worlds to answer query Pr(*a*) or given evidence $\Pr_e(a)$
- How to avoid these problems?
  - No general solution
  - Exploit structure
- Use conditional independence

Independence

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

  Pr(*toothache*,*catch*,*cavity*,*weather*) = Pr(*toothache*,*catch*,*cavity*)Pr(*weather*)

- 32 entries reduced to 12; for *n* independent biased coins, $2^n \rightarrow n$
- Absolute independence powerful but rare
  - Also the first assumption to try in machine learning
- Dentistry is a large field with hundreds of variables none of which are independent. What do we do?

Conditional Independence

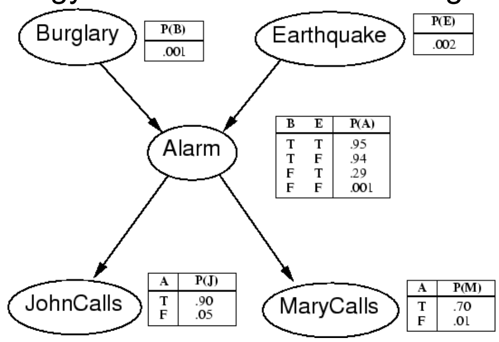
- Pr(*toothache*,*cavity*,*catch*) has $2^3 - 1 = 7$ independent entries
- If I have a cavity, the probability that the probe catches in it doesn't depend on whether I have a toothache:
  Pr(*catch*|*toothache*,*cavity*) = Pr(*catch*|*cavity*)
- The same independence holds if I haven't got a cavity:
  Pr(*catch*|*toothache*,¬*cavity*) = Pr(*catch*|¬*cavity*)
- *Catch* is conditionally independent of *toothache* given *cavity*:
  Pr(*catch*|*toothache*,*cavity*) = Pr(*catch*|*cavity*)
- Equivalent statements:
  Pr(*toothache*|*catch*,*cavity*) = Pr(*toothache*|*cavity*)
  Pr(*toothache*,*catch*|*cavity*) = Pr(*toothache*|*cavity*)Pr(*catch*|*cavity*)

Belief Networks

- A belief network (Bayes net) represents the dependence between variables
- Components of a belief network graph:
  - Nodes
    - These represent variables
  - Links
    - X points to Y if X has a direct influence on Y
  - Conditional probability tables
    - Each node has a CPT that quantifies the effects the parents have on the node
- The graph has no directed cycles!!!

Example

- I'm at work, neighbor John calls to say my alarm is ringing, but neighbor Mary doesn't call. Sometimes the alarm is set off by minor earthquakes. Is there a burglar?
- Variables: *Burglary*, *Earthquake*, *Alarm*, *JohnCalls*, *MaryCalls*
- Network topology reflects "causal" knowledge:



Semantics of a Bayes Network

What does this actually mean?

- The polytree structure of a BN means: every X_i is conditionally independent of all of its nondescendants given its parents:

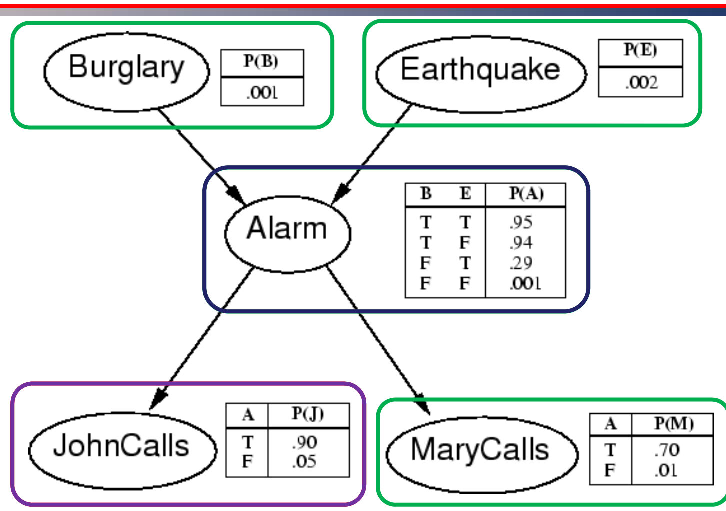
$$\Pr(X_i \mid S \cup Parent(X_i)) = \Pr(X_i \mid Parent(X_i))$$

$$\text{for any subset } S \subseteq \text{NonDescendant}(X_i)$$

- If we ask for $\Pr(x_1, x_2, \ldots, x_n)$ and we have an ordering consistent with the network.
- By the chain rule:
  - $\Pr(x_1, x_2, \ldots, x_n) = \Pr(x_n \mid x_{n-1}, \ldots, x_1)\Pr(x_{n-1} \mid x_{n-2}, \ldots x_1)\ldots\Pr(x_1)$
  - $\Pr(x_n \mid Parent(x_{n-1}))\Pr(x_{n-1} \mid Parent(x_{n-2}))\ldots\Pr(x_1)$

This works, because all of $x_1, \ldots, x_n$ are CONDITIONALLY INDEPENDENT from each other

…is equal to the probability that JohnCalls, given Alarm.



Constructing a Bayes Network

- Given any distribution over variables, a BN can be generated to represent the distribution
- BN's are generally generated by hand.
- The ordering of the variable set can make a difference in the BN
  - The more the ordering reflects causal intuitions, the smaller the BN (variables parents only come earlier in the ordering)

Constructing a Bayes Network

- 1. Choose an ordering of variables $X_1,..X_n$
- 2. For i = 1 to n
  - Add X_i to the network — i.e. add it as a node
  - Select parents from $X_1,..X_{i-1}$ such that $Pr(X_i|Parents(X_i))=Pr(X_i|X_1,\dots,X_{i-1})$ — i.e. add links from nodes already in the graph
- This choice of parents guarantees the global semantics:

$$\Pr(X_1,\dots,X_n) = \prod_{i=1}^{n} \Pr(X_i \mid X_1,\dots X_{i-1}) \quad \text{chain rule}$$

$$= \prod_{i=1}^{n} \Pr(X_i \mid Parents(X_i)) \quad \text{by construction}$$

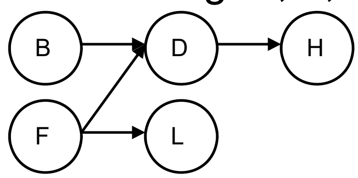
Example

- Suppose you are going home, and you want to know the probability that the lights are on given the dog is barking and the dog does not have a bowel problem. If the family is out, often the lights are on. The dog is usually in the yard when the family is out and when it has bowel troubles. If the dog is in the yard, it probably barks.

- Use the variables:
  F = family out
  L = light on
  B = bowel problem
  D = dog out
  H = hear bark

Example

- So choose an ordering: F, L, B, D, H

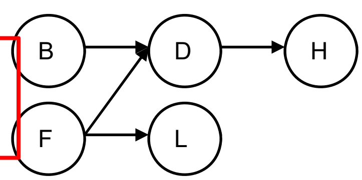


- Pr(L|F) = Pr(L)?
- Pr(B|L,F) = Pr(B|L)? Pr(B|F,L) = Pr(B)?
- Pr(D|F,B,L) = Pr(D|F,B)? Yes
- Pr(H|F,L,B,D) = Pr(H|D)? Yes

Example

- We know:
  - L is directly influenced by F and is independent of B, D, H given F
    Add link from F to L
  - D is directly influenced by F and B, independent of L, H given F,B
    Add link from F to D and B to D
  - H is directly influenced by D and is independent of F, L, B given D
    Add link from D to H

This Bayes Network says:
Pr(F,L,B,D,H) =
Pr(F)Pr(B)Pr(D|F,B)Pr(L|F)Pr(H|D)



- Once we specify the topology, we need to specify the conditional probability table for each node.

Pr(f) = 0.15, 0.85 Pr(b) = 0.01, 0.99
Pr(l|f) = 0.60, 0.40 Pr(l|¬f) = 0.05, 0.95
Pr(d|f,b) = 0.99, 0.01 Pr(d|f,¬b) = 0.90, 0.10
Pr(d|¬f,b) = 0.97, 0.03 Pr(d|¬f,¬b) = 0.30, 0.70
Pr(h|d) = 0.70, 0.30 Pr(h|¬d) = 0.01, 0.99

Independence Review

- Variables x and y are <u>independent</u> iff:
  - Pr(x) = Pr(x|y) iff Pr(y) = Pr(y|x) iff Pr(xy) = Pr(x)Pr(y)
  - Learning about y doesn't influence beliefs about x
- x and y are <u>conditionally independent given</u> z iff:
  - Pr(x|z) = Pr(x|yz) iff Pr(y|z) = Pr(y|xz) iff Pr(xy|z) = Pr(x|z)Pr(y|z) iff…
  - Learning y doesn't influence your beliefs about x **if you already knew z**.
  - Learning your grade on an exam can influence the probability of getting an A for AI; but if you already knew your final AI grade, learning the exam grade wouldn't influence your grade assessment.

Variable Independence

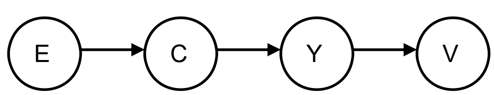
- Two variables X and Y are conditionally independent given Z iff x,y are conditionally independent given z for all $x \in Dom(X)$, $y \in Dom(Y)$, $z \in Dom(Z)$
  - Also applies to sets of variables **X**,**Y**,**Z**
- If you know the value of Z nothing you learn about Y will influence your beliefs about X
- Also, each node is conditionally independent given its Markov blanket: parents + children + children's parents

What effect does Independence have?

- If $X_1, X_2, \ldots X_n$ are mutually independent
  - A full joint distribution requires only *n* parameters instead of 2^{n-1}.
- Unfortunately complete mutual independence is rare. Most realistic domains do not have this property.
- Fortunately, most domains do exhibit some conditional independence. Bayes networks represent this.

Exploiting Conditional Independence

- Example:
  - If I wake too early (E), I will be crabby (C). If I am crabby, there is a chance I will yell at someone (Y). If Y, there is an increased chance I will lose my voice (V).



- If you learn any of E, C, or Y, your assessment of Pr(V) will change.
  - Pr(V) is not independent of E, C, and Y

Exploiting Conditional Independence

$$E \rightarrow C \rightarrow Y \rightarrow V$$

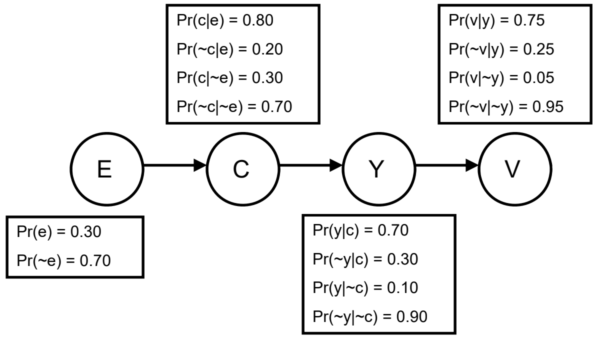
- But if you knew Y (true or false), learning values of E or C would not influence Pr(V). The influence of E and C is mediated by the influence of Y.
  - I don't lose my voice because I woke early but because I yell.
  - So V is independent of E and C given Y.

Exploiting Conditional Independence

$$E \rightarrow C \rightarrow Y \rightarrow V$$

- This means:
  - $Pr(V|Y, \{E,C\}) = Pr(V|Y)$
  - $Pr(Y|C,\{E\}) = Pr(Y|C)$
  - Pr(C|E) and Pr(E) doesn't simplify
- By the chain rule
  - $Pr(V,Y,C,E) = Pr(V|Y,C,E)Pr(Y|C,E)Pr(C|E)Pr(E)$
- By our independence assumptions:
  - $Pr(V,Y,C,E) = Pr(V|Y)Pr(Y|C)Pr(C|E)Pr(E)$
  - The full joint probability can be specified with 4 local conditional distributions.

Example Quantification



- Specifying the joint requires only 7 parameters (note that half of these are "1 minus" the others).
  - Linear in number of variables if dependence has a chain structure.

Inference

- Inference of $\Pr(\alpha)$ is simply summing out rule:

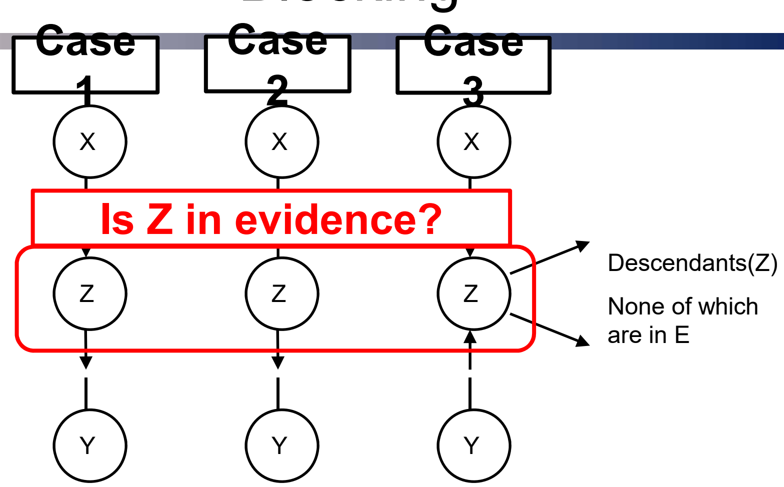
$$\Pr(\alpha) = \sum_{c_i \in Dom(C)} \Pr(\alpha \mid c_i)\Pr(c_i)$$
$$= \sum_{c_i \in Dom(C)} \Pr(\alpha \mid c_i) \sum_{e_i \in Dom(E)} \Pr(c_i \mid e_i)\Pr(e_i)$$

- Computing:
  - Pr(c) = Pr(c|e)Pr(e) + Pr(c|¬e)Pr(¬e) = 0.80*0.30+0.3*0.70 = 0.45
  - Pr(¬c) = Pr(¬c|e)Pr(e) + Pr(¬c|¬e)Pr(¬e) = 0.20*0.30+0.70*0.70 = 0.55
  - Pr(¬y) = Pr(¬y|c)Pr(c) + Pr(¬y|¬c)Pr(¬c) = 0.30*0.45 +0.90*0.55 = 0.63
  - Pr(y) = 1 – Pr(¬y) = 0.37

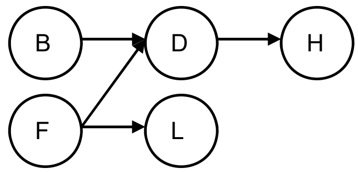
D-Separation

- Given a BN, we can determine if two variables X, Y are independent using D-separation
  - A set of variables E <u>d-separates</u> X and Y if it <u>blocks</u> every undirected path in the BN between X and Y
  - X and Y are conditionally independent given E if E d-separates X and Y
- If path relation P is an undirected path from X to Y with evidence set E. We say E blocks path P iff there is some node Z on the path such that:
  - Case 1: one arc on P goes into Z and one goes out of Z, and $Z \in E$
  - Case 2: both arcs on P leave Z, and $Z \in E$
  - Case 3: both arcs on P enter Z and neither Z, nor any of its descendants are in E.

Blocking



D-Separation: Intuitions

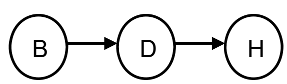


- B and H are dependent, but independent given D, since D blocks the path (Case 1).
- L and D are dependent, but are independent given F since F blocks the path (Case 2).
- F and B are independent, D blocks the path, since it is not in evidence nor is its descendant H (Case 3).

Simple Forward Inference

- Computing prior probabilities requires simple forward "propagation" of probabilities.

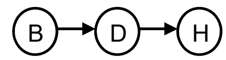
$$\Pr(H) = \sum_{D,B} \Pr(H \mid D, B)\Pr(D, B)$$
$$= \sum_{D,B} \Pr(H \mid D)\Pr(D \mid B)\Pr(B)$$
$$= \sum_{D} \Pr(H \mid D)\sum_{B} \Pr(D \mid B)\Pr(B)$$



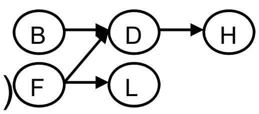
- (1) follows by summing out rule; (2) by chain rule and independence; (3) by distribution of sum
  - Note: all (final) terms are CPT's in the BN
  - Note: only ancestors of D considered

Simple Forward Inference (Chain)

- Same idea applies when we have "upstream" evidence
  - $\Pr(H|b) = \sum_D \Pr(H|D,b)\Pr(D,b)$
    $= \sum_D \Pr(H|D)\Pr(D|b)$



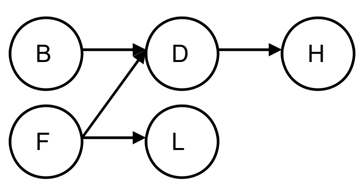
- Same idea applies with multiple parents (Pooling)
  - $\Pr(D) = \sum_{F,B} \Pr(D|F,B)\Pr(F,B)$
    $= \sum_{F,B} \Pr(D|F,B)\Pr(F)\Pr(B)$



  - (1) follows by summing out rule; (2) by independence of F,B

Simple Forward Inference (Pooling)

- Same idea with evidence:
  - $\Pr(D|f,b) = \Pr(D|f,b)\Pr(f,b)$
    $= \Pr(D|f,b)\Pr(f)\Pr(b)$



Simple Backward Inference

- When evidence is downstream of query variable, we must reason "backwards." This requires the use of Bayes rule:
  - $\Pr(B|h) = \alpha \Pr(h|B)\Pr(B)$
    $= \alpha \sum_{DB} \Pr(h|D,B)\Pr(D|B)\Pr(B)$
    $= \alpha \sum_{D} \Pr(h|D) \sum_{B} \Pr(D|B)\Pr(B)$

    $B \rightarrow D \rightarrow H$
- First step is just Bayes rule
  - Normalizing constant α is 1/Pr(h); but we don't need to compute it explicitly if we compute Pr(B|h) for each value of B; we just add up the terms Pr(h|B)Pr(B) for all values of B (they sum to Pr(h)).

Backward Inference

- Same idea applies when several pieces of evidence lie "downstream"
  - Same steps as before; but now we compute probability of both pieces of evidence given hypothesis and combine them.
  - Note: simplification down to CPTs will require finding independence.

Variable Elimination

- The above examples give us a simple inference process for networks without loops.
- The process can be improved by eliminating repeated calculations.
- The variable elimination algorithm is a dynamic programming algorithm that applies the summing out rule repeatedly, exploiting the independence of the network and the ability to distribute sums inward.

Factor

- A function $f(X_1,X_2,\ldots,X_n)$ is called a factor.
- A tabular representation of a factor is exponential in n (just like a joint probability distribution).
- Each CPT in a BN is a factor:
  - Pr(C|A,B) is a factor of three variables A, B, C
  - Notation: $f(\mathbf{X},\mathbf{Y})$ denotes a factor over the variable sets **X** and **Y**.

The Product of Two Factors

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

- Let *f*(**X**,**Y**) and *g*(**Y**,**Z**) be two factors with **Y** in common.
- The product of *f* and *g*, *h* = *f* x *g* is:

h(**X**,**Y**,**Z**) = *f*(**X**,**Y**) x *g*(**Y**,**Z**)

f(**X**,**Y**)		*g*(**Y**,**Z**)		*h*(**X**,**Y**,**Z**)			
xy	0.9	yz	0.7	xyz	0.63	xy¬z	0.27
x¬y	0.1	y¬z	0.3	x¬yz	0.8	x¬y¬z	0.02
¬xy	0.4	¬yz	0.8	¬xyz	0.28	¬xy¬z	0.12
¬x¬y	0.6	¬y¬z	0.2	¬x¬yz	0.48	¬x¬y¬z	0.12

Summing a Variable Out of a Factor

$$\Pr(a)=\sum_{b\in Dom(B)}\Pr(a \mid b)\Pr(b)$$

- Let *f*(**X**,**Y**) be a factor with variable **X**.
- We sum out each variable $x \in$Dom(**X**) from *f* to produce a new factor *h*= $\sum_X f$, which is:

h(**Y**) = $\sum_{x\in \mathrm{Dom}(X)}$ *f*(**X**,**Y**)

f(**X**,**Y**)		*h*(**Y**)	
xy	0.9	y	1.3=0.65
x¬y	0.1	¬y	0.7=0.35
¬xy	0.4		
¬x¬y	0.6		

These two need to add to 1

Restricting a Factor

$$\Pr(a) = \Pr(a \mid b)$$

- Let $f(\mathbf{X},\mathbf{Y})$ be a factor with variable $\mathbf{X}$.
- We restrict factor f to $X=x$ by setting X to the value x and "deleting". Define $h = f_{X=x}$ as:

$$h(\mathbf{Y}) = f(x,\mathbf{Y})$$

$f(\mathbf{X},\mathbf{Y})$		$h(\mathbf{Y}) = F_{\mathbf{X}=x}$	
xy	0.9	y	0.9
x¬y	0.1	¬y	0.1
¬xy	0.4		
¬x¬y	0.6		

Variable Elimination Algorithm

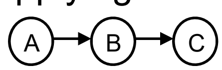
- Given query variable Q, and remaining variables sets $\mathbf{Z}$. Let F be the set of factors corresponding to CPTs for $\{Q\} \cap \mathbf{Z}$
  1. Choose an elimination ordering $Z_1,\ldots Z_n$ of variables in Z.
  2. For each Z_j, in the order given, eliminate $Z_j \in \mathbf{Z}$ as follows:
     1. Compute new factor $g_j = \sum_{Zj} f_1 x f_2 x \ldots x f_k$, where the f_i are the factors in F that include Z_j
     2. Remove the factors f_i (that mention Z_j) from F and add new factor g_j to F
  3. The remaining factors refer only to the query variable Q. Take their product and normalize to produce $\Pr(Q)$.

Variable Elimination

- One way to think of variable elimination:
  - Write out desired computation using the chain rule, exploiting the independence relations in the network
  - Arrange the terms in a convenient fashion
  - Distribute each sum (over each variable) in as far as it will go
    - i.e., the sum over variable *X* can be "pushed in" as far as the "first" factor mentioning *X*
  - Apply operations "inside out", repeatedly eliminating and creating new factors (note that each step/removal of a sum eliminates one variable)
- Variable Elimination explained by Peter Norvig
  - YouTube: Unit 4, 4 Variable Elimination
  - This is a series of six videos (a-f)

Variable Elimination: No Evidence

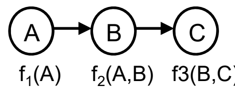
- Computing prior probabilities of query variable *X* can be seen as applying the operation of factors



$$
\begin{aligned}
Pr(C) &= \sum_{A,B} Pr(C|B)Pr(B|A)Pr(A) \\
&= \sum_B Pr(C|B)\sum_A Pr(B|A)Pr(A) \\
&= \sum_B f_3(B,C)\sum_A f_2(A,B)f_1(A) \\
&= \sum_B f_3(B,C)f_4(B) \\
&= f_5(C)
\end{aligned}
$$

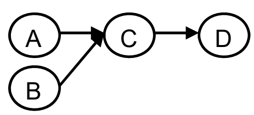
Define new factors: $f_4(B) = \sum_A f_2(A,B)f_1(A)$ and $f_5(C) = \sum_B f_3(B,C)f_4(B)$

Variable Elimination: No Evidence



$f_1(A)$		$f_2(A,B)$		$f_3(B,C)$		$f_4(B)$		$f_5(C)$	
a	0.9	ab	0.9	bc	0.7	b	0.85	c	0.625
¬a	0.1	a¬b	0.1	b¬c	0.3	¬b	0.15	¬c	0.375
		¬ab	0.4	¬bc	0.2				
		¬a¬b	0.6	¬b¬c	0.8				

VE: No Evidence Example 2



$$
\begin{aligned}
Pr(D) &= \sum_{A,B,C} Pr(D|C)Pr(C|B,A)Pr(B)P(A) \\
&= \sum_C Pr(D|C)\sum_B Pr(B)\sum_A Pr(C|B,A)Pr(A) \\
&= \sum_C f_4(D,C)\sum_B f_2(B)\sum_A f_3(C,B,A)f_1(A) \\
&= \sum_C f_4(D,C)\sum_B f_2(B)f_5(B,C) \\
&= \sum_C f_4(D,C)f_6(C) \\
&= f_7(D)
\end{aligned}
$$

Define new factors: $f_5(B,C) = \sum_A f_3(C,B,A)f_1(A)$ and $f_6(C) = \sum_B f_2(B)f_5(B,C)$ and $f_7(D) = \sum_C f_4(D,C)f_6(C)$

Variable Elimination with Evidence

- Computing posterior of query variable given evidence is similar:

$$A \rightarrow B \rightarrow C$$

$$\begin{aligned}
\Pr(A|c) &= \alpha \Pr(A)\Pr(c|A) \\
&= \alpha \Pr(A)\textstyle\sum_B \Pr(c|B)\Pr(B|A) \\
&= \alpha f_1(A)\textstyle\sum_B f_3(B,c)f_2(A,B) \\
&= \alpha f_1(A)\textstyle\sum_B f_4(B)f_2(A,B) \\
&= \alpha f_1(A)f_5(A) \\
&= \alpha f_6(A)
\end{aligned}$$

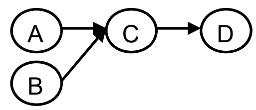
Define new factors: $f_4(B) = f_3(B,c)$ and $f_5(A) = \sum_B f_4(B)f_2(A,B)$ and $f_6(A) = f_1(A)f_5(A)$.

Variable Elimination with Evidence

- Given query variable *Q*, evidence variable sets **E** (observed to be **e**), remaining variable sets **Z**. Let *F* be the set of factors involving CPTs for {Q} $\cap$ **Z**.

1. Replace each factor $f \in F$ that mentions a variable(s) in **E** with its restrictions $f_{\mathbf{E}=\mathbf{e}}$
2. Choose an elimination ordering $Z_1, \ldots, Z_n$ of variables in *Z*.
3. Run variable elimination as above.
4. The remaining factors refer only to the query variable *Q*. Take their product and normalize to produce Pr(*Q*).

VE Example 2 with Evidence



$Pr(A|d) = \alpha Pr(A)Pr(d|A)$

$= \alpha Pr(A)\sum_B Pr(d|C)Pr(C|A,B)Pr(B)$

$= \alpha f_1(A)\sum_B f_2(B)\sum_C f_4(C,d)f_3(A,B,C)$

$= \alpha f_1(A)\sum_B f_2(B)\sum_C f_5(C)f_3(A,B,C)$

$= \alpha f_1(A)\sum_B f_2(B)f_6(A,B)$

$= \alpha f_1(A)f_7(A)$

Last factors $f_7(A)$,$f_1(A)$ The product $f_1(A) \times f_7(A)$ is (possibly unnormalized) posterior. So… $Pr(A|d) = \alpha f_1(A) \times f_7(A)$

Some Notes on the VE Algorithm

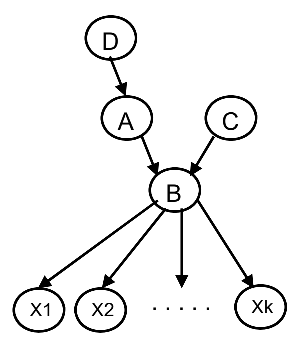
- After iteration *j* (elimination of Z_j), factors remaining in set *F* refer only to variables $X_{j+1},\ldots,Z_n$ and *Q*. No factor mentions an evidence variable **E** after the initial restriction.
- Number of iterations: linear in number of variables
- Complexity is linear in number of variables and exponential in size of the largest factor.
  - Recall each factor has exponential size in its number of variables).
  - Can't do any better than size of BN, since its original factors are part of the factor set.
  - When we create new factors, we might make a set of variables larger.

Some Notes Continued

- The size of the resulting factors is determined by elimination ordering
- For polytrees, easy to find a good ordering (outside to inside)
- For general BNs, sometimes good orderings exist, sometimes they don't
  - Finding the optimal ordering for a general BN is NP-hard.
  - Inference in a general BN is NP-hard.

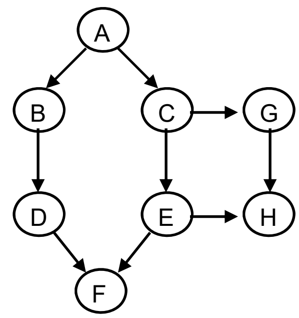
Elimination Ordering: Polytrees

- Inference is linear in size of network
  - Ordering: eliminate only "singly-connected" nodes
    - Eliminate D,C,X_1,…X_k, A, B
    - Result is no factor larger than original CPTs
    - Eliminating B before these gives factors include all of A, X_1,..,X_k

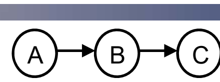


Effects of Different Orderings

- Suppose query variable is D. Consider different orderings for this network
  - A,F,H,G,B,C,E:
  - E,C,A,B,G,H,F:
- Which ordering creates smallest factors?



Relevance



- Certain variables have no impact on the query. In the ABC network, computing Pr(A) with no evidence requires elimination of B and C.
  - But when you sum out these variables, you compute a trivial factor
  - Eliminating C: $f_4(B) = \sum_C f_3(B,C) = \sum_C Pr(C|B)$
  - 1 for any value of B: $(Pr(c|b)+Pr(\neg c|b)=1$

Existing Systems

- JavaBayes
  - http://www-2.cs.cmu.edu/~javabayes/Home/
- Bayes Software List
  - https://www.cs.ubc.ca/~murphyk/Software/bnsoft.html

  - Create a JavaBayes network with 5 nodes and 4 links, as indicated in our example. Use the CPT values we have specified.
  - Calculate posterior probabilities by selecting Observe then the observed value for each observed variable
  - Next, query a variable by selecting Query then the node
  - For example, calculate $P(l|h,b_c)$, the posterior probability of LightOn given HearBark and not BowelProblem
  - Query LightOn, generates rules

The Bad (and Challenging) News

- General querying of Bayes nets is NP-hard
- The best known algorithm is exponential in the number of variables
- Pathfinder system
  - Heckerman, 1991
  - Diagnostic system for lymph-node diseases
  - 60 diseases, 100 symptoms and test rulests
  - 14,000 probabilities
  - 8 hours to determine variables, 35 hours for topology, 40 hours for CPTs
  - Outperforms world experts in diagnosis
  - Being extended to several dozen other medical domains

Inference by Stochastic Simulation

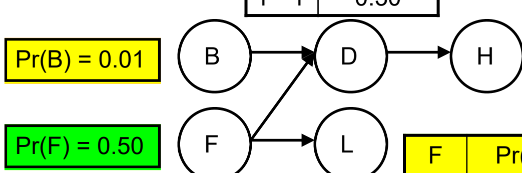
- Basic idea:
  1. Draw N samples from a sampling distribution S
  2. Compute an approximate posterior probability P'
  3. Show this converges to the true probability P

- 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

Empty Network Example

- For n events
  - Spin a roulette wheel at each node biased by the CPT

F	B	Pr(D)
t	t	0.99
t	f	0.90
f	t	0.97
f	f	0.30



D	Pr(H)
t	0.70
f	0.01

F	Pr(L)
t	0.60
f	0.05

Rejection Sampling

- Provides a method to compute conditional probabilities Pr(X|**E**)
  1. Perform empty network sampling
  2. Collect those samples that match the evidence
  3. Compute the probability of the query

- Estimate the probability the dog is out given the light is on, Pr(D|l)
  - Over 100 samples, the light is on for 64
  - Of these 64, the dog is out 41 of the samples
  - Pr(D|l) = $\alpha\langle 41,23\rangle = \langle 0.641,0.359\rangle$

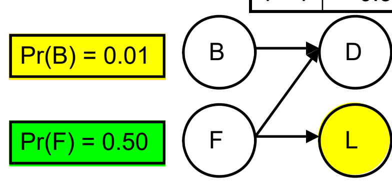
Likelihood Weighting

- Rejection sampling is inefficient, why not just sample based on the evidence
  1. Fix evidence variables
  2. Sample only nonevidence variables
  3. Weight each sample by the likelihood it accords the evidence
- The weighting makes up for the difference between the actual and desired sampling distributions.

Likelihood Weighting Example

What is the probability that the dog has bowel problems given we heard it bark, and the light is on?
Pr(B|l,h)

F	B	Pr(D)
t	t	0.99
t	f	0.90
f	t	0.97
f	f	0.30



D	Pr(H)
t	0.70
f	0.01

F	Pr(L)
t	0.60
f	0.05

w = 1.0 * 0.60 * 0.70 = 0.42

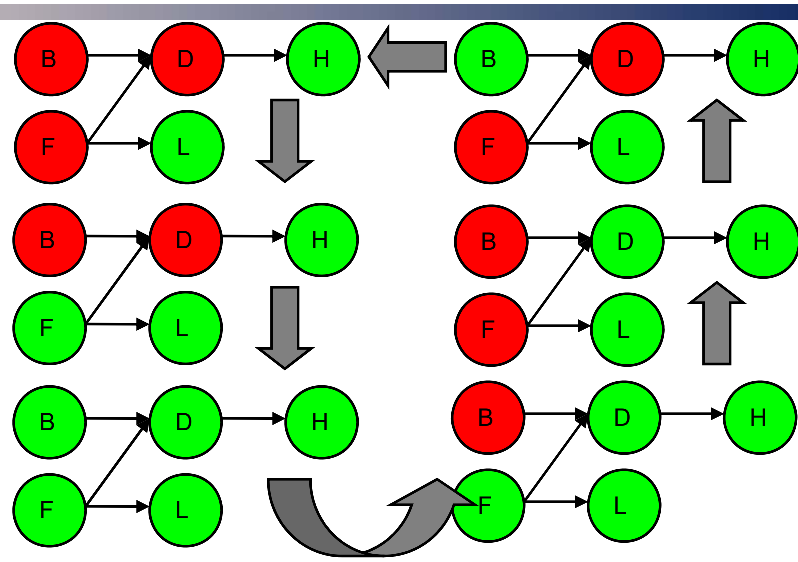
Pr(B=false|L=true, H=true) = 0.42

Approximate Inference using Markov chain Monte Carlo (MCMC)

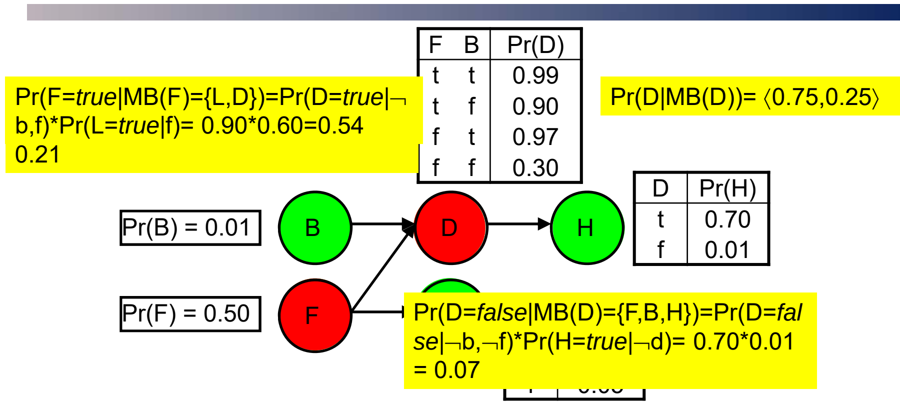
- Represents the network as a 'state' – the current assignments to all variables
- Generate next state by sampling one variable given the Markov blanket
  - This variable choice can be made at random
- Sample each variable in turn, keeping evidence fixed

$$\Pr(x_i' \mid MB(X_i)) = \alpha \Pr(x_i' \mid Parents(X_i)) \prod_{z_j \in Children(X_i)} \Pr(z_j \mid Parents(Z_j))$$

The Markov chain



Markov Chain Example



F	B	Pr(D)
t	t	0.99
t	f	0.90
f	t	0.97
f	f	0.30

D	Pr(H)
t	0.70
f	0.01

Randomly assign true/false values to nodes
Until no probability change
Pick a node, and update it's probability

Bayesian Network Evaluation Summary

- Exact inference by variable elimination:
  - Polytime on polytrees, NP-hard on general graphs
  - Space = time, very sensitve 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

Dempster-Shafer Theory

- Measure certainty
- Belief(∅) = 0
  ∅ is the null set
- Belief(⊖) = 1 ⇒ Bel(H)+Bel(¬H)+Bel(⊖) = 1.0
  ⊖ is the set of all possible outcomes
- For every non-negative integer n and every $\{A_i | i=1,2,\ldots,n\} \subseteq \ominus$

$$Bel(A_i) \geq \sum_{I \subset \{A_1,\ldots,A_n\}; I \neq null} -1^{I+1} Bel\left(\bigcap_{i=I} A_i\right)$$

- Facts and rules have beliefs, propagate belief values
- Represents certainty about certainty

Fuzzy Logic

- Fuzzy Logic is a multivalued logic that allows intermediate values to be defined between conventional evaluations like yes/no, true/false, etc.
- Fuzzy Logic was initiated in 1965 by Lotfi A. Zadeh, professor of computer science at the University of California in Berkeley.
- The concept of fuzzy sets is associated with the term “graded membership”.
- This has been used as a model for inexact, vague statements about the elements of an ordinary set.
- Fuzzy logic prevalent in products such as:
  - Washing machines
  - Video cameras
  - Razors
  - Subway systems

Fuzzy Sets

- In a fuzzy set the elements have a DEGREE of existence.
- Some typically fuzzy sets are “large numbers”, “tall men”, “young children”, “approximately equal to 10”, “mountains”, etc.

Ordinary Sets

$$f_A(x) = \begin{cases} 1 & \text{if } x \text{ in } A \\ 0 & \text{if } x \text{ not in } A \end{cases}$$

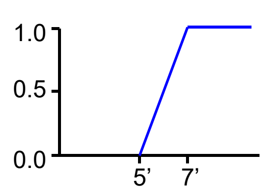
- P(X) is the "power set" of X (all subsets of X)
- 2^X represents all functions from X into {0,1}
- $f_{A \cap B} = \max(f_A, f_B)$

Fuzzy Sets

- $f_A(x) = i$, where $0 \leq i \leq 1$
- if $f_A(x) > f_A(y)$, then x is "more in" the set A than y
- if $f_A(x) = 1$, then $x \in A$ if $f_A(x) = 0$, then $x \notin A$ if $f_A(x) = \lambda$, where $0 \leq \lambda \leq 1$, then $x \in_\lambda A$

- Degree of membership sometimes determined as a function (degree of tall calculated as a function of height)

$$tall(x) = \begin{cases} 0 & \text{if } height(x) < 5' \\ \dfrac{height(x) - 5'}{2'} & \text{if } 5' \leq height(x) \leq 7' \\ 1 & \text{if } height(x) > 7' \end{cases}$$



Fuzzy Set Relations

- One set A is a "subset" of set B if for every x, $f_A(x) \le f_B(x)$
- Sets A and B are equal if for every element x, $f_A(x) = f_B(x)$.
- OR / Union: $A \cup B$ is the smallest fuzzy subset of X containing both A and B, and is defined by $f_{A \cup B}(x) = max(f_A(x), f_B(x))$
- AND / Intersection: The intersection $A \cap B$ is the largest fuzzy subset of X contained in both A and B, and is defined by $f_{A \cap B}(x) = min(f_A(x), f_B(x))$
- NOT: truth($\neg$x) = 1.0 - truth(x)
- IMPLICATION: $A \rightarrow B \equiv \neg A \vee B$, so truth($A \rightarrow B$) = $min(f_A(x), f_B(x))$

Fuzzy Example

- Fuzzy Inverted Pendulum Controller
  - http://rorchard.github.io/FuzzyJ/FuzzyPendulum.html

Review

- Uncertainty Reasoning
- Next Class:
  - Decision and Game Theory

Decision in the Face of Uncertainty

- Preferences
- Utility/Value
- Maximum Expected Utility (MEU/MEV)
- Syntax and Semantics
- Inference
- Independence and Bayes' Rule

Utility/Value Theory

- Suppose I believe the following:
  $Pr(A_{25}| ..) = 0.04$
  $Pr(A_{90}| ..) = 0.70$
  $Pr(A_{120}| ..) = 0.95$
  $Pr(A_{1440}| ..) = 0.99$

- Which action to choose?
  - Probability theory tells us which is most likely
  - But depends on my preferences for missing flights vs. airport food, etc.
    - Utility theory is used to represent and infer preferences
    - Decision theory = utility theory + probability theory

Preference Orderings

- A preference ordering $\succcurlyeq$, $\succ$ is a ranking of all possible states (S).
  - These could be outcomes of actions, truth assets, states in a search problem, etc.
  - $s \succcurlyeq t$ means that state s is at least as good as t
  - $s \succ t$ means that state s is strictly preferred to t
- We insist that $\succcurlyeq$ is:
  - Reflexive: $s \succcurlyeq s$ for all states s
  - Transitive: if $s \succcurlyeq t$ and $t \succcurlyeq w$ then $s \succcurlyeq w$ ($w \succcurlyeq s$)
  - Connected: for all states s, t, either $s \succcurlyeq t$ or $t \succcurlyeq s$

Why Impose These Conditions?

- Structure of preference ordering imposes "rationality requirements" (a partial ordering)
- A <u>decision problem under certainty</u> is:
  - A set of <u>decisions</u> D
    - Paths in a search graph, plans, actions, etc.
  - A set of <u>outcomes</u> or states S
    - States in a plan, etc.
  - An <u>outcome</u> <u>function</u> f: $D \rightarrow S$
    - The outcome of the decision
  - A preference ordering $\succcurlyeq$ over S
- A <u>solution</u> to a decision problems is any $d \in D$ s.t. $f(d) \succcurlyeq f(d')$ for all $d' \in D$

Decision Making under Uncertainty

- Suppose actions don't have deterministic outcomes
  - When our robot gets coffee, it spills 20% of the time
- What should an agent do?
  - Decision to 'get coffee' leads to good and bad outcomes
  - Doing nothing leads to a mediocre outcome
- Should agent be optimistic? Pessimistic?
- The odds of success should influence the decision

Utilities

- Rather than just ranking outcomes, we must quantify our degree of preference
  - How much more important is getting coffee over not having a mess
- A set of utility functions U: $S \rightarrow R$ associates a real valued utility with each outcome
  - $U(s)$ measures your degree of preference for s
  - Note: U induces a preference ordering $\succcurlyeq_U$ over S defined as: $s \succcurlyeq_U t$ iff $U(s) \geq U(t)$

Expected Utility

- Under conditions of uncertainty, each decision d induces a distribution $\Pr_d$ over possible outcomes
  - $\Pr_d(s)$ is probability of outcome s under decision d
- The expected utility of decision d is:

$$EU(d) = \sum_{s \in S} \Pr_d(s) U(s)$$

- Given c = have coffee and m = massive spill
- If $U(c,\neg m) = 10$, $U(\neg c,\neg m) = 5$, $U(\neg c,m) = 0$ then $EU(\text{getcoffee}) = 8$ and $EU(\text{donothing}) = 5$
- If $U(c,\neg m) = 10$, $U(\neg c,\neg m) = 9$, $U(\neg c, m) = 0$ then $EU(\text{getcoffee}) = 8$ and $EU(\text{donothing}) = 9$

Maximum Expected Utility

- The principle of maximum expected utility (MEU) states that the optimal decision under conditions of uncertainty is to choose the action leading to the greatest expected utility
- In our example
  - If my utility functions is the first one, my robot should get coffee
  - If your utility function is the second one, your robot should do nothing

Decision Problems: Uncertainty

- A decision problem under uncertainty is:
  - A set of decision *D*
  - A set of outcomes or states *S*
  - An outcome function Pr: $D \rightarrow \Delta(S)$
    - $\Delta(S)$ is the set of distributions over S ($\Pr_d$)
  - A utility function *U* over *S*
- A solution to a decision problem under uncertainty is any $d \in D$ s.t. $EU(d) \succcurlyeq EU(d')$ for all $d' \in D$

Expected Utility: Notes

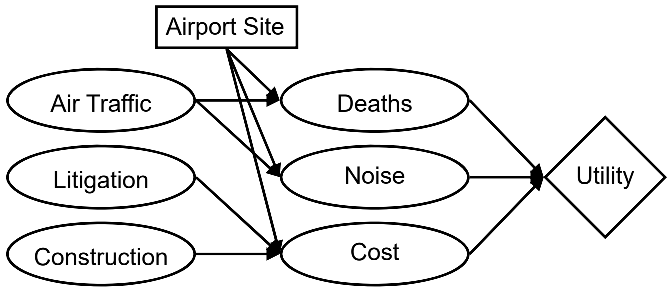
- Note that this viewpoint accounts for:
  - Uncertainty in action outcomes
    - Getting coffee with our robot
  - Uncertainty in state of knowledge
    - If I need to get in an office will it be unlocked when I get there?
  - Any combination of the two
    - Will the office be unlocked so I can deliver coffee?

Expected Utility: Notes

- Why MEU/ Where do utilities come from?
  - Underlying foundations of utility theory tightly couple utility with action/choice
  - A utility function can be determined by asking someone about their preferences for actions in specific scenarios ("lotteries" over outcomes)
- Utility functions needn't be unique
  - If I multiply U by a positive constant, all decisions have same relative utility
  - If I add a constant to U, same thing
  - U is unique up to positive affine transformation

Decision Networks

- Add action nodes and utility nodes to belief networks to enable rational decision making



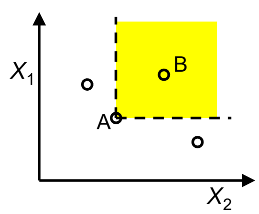
- Algorithm:
  For each value of action node
  compute expected value of utility node given action, evidence
  Return MEU action

Multiattribute Utility

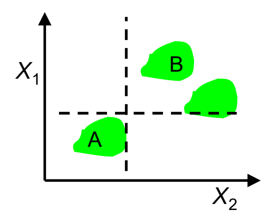
- How can we handle utility functions of many variables $X_1,\ldots,X_n$? i.e. U(Deaths, Noise, Cost)?
- How can complex utility functions be assessed from preference behavior?
- Idea 1: Identify conditions under which decision can be made without complete identification of $U(x_1,\ldots,x_n)$
- Idea 2: Identify various types of independence in preferences and derive consequent canonical forms for $U(x_1,\ldots,x_n)$

Strict Dominance

- Solution to idea 1: Typically define attributes such that U is monotonic in each
- Strict dominance: choice B strictly dominate choice A iff $\forall i$ $X_i(B) \geq X_i(A)$ (and hence $U(B) \geq U(A)$)

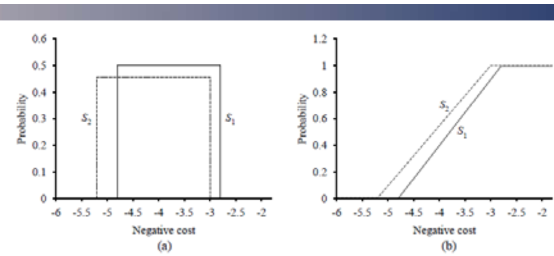


Deterministic Attributes



Uncertain Attributes

Stochastic Dominance



- Distribution p_1 stochastically dominates distribution p_2 iff

$$\forall x \int_{-\infty}^{x} p_1(x')dx' \leq \int_{-\infty}^{x} p_2(x')dx'$$

- If U is monotonic in x, then A_1 with outcome distribution p_1 stochastically dominates A_2 with outcome distribution p_2:

$$\int_{-\infty}^{\infty} p_1(x)U(x)dx \leq \int_{-\infty}^{\infty} p_2(x)U(x)dx$$

- Multiattribute case: stochastic dominance on all attributes $\Rightarrow$ optimal

Preference Structure: Deterministic

- As with Bayes nets, want to identify regularities in the utility preference structure using representation theorems
  $U(x_1,\ldots x_n) = f[f_1(x_1),\ldots,f_n(x_n)]$
- X_1 and X_2 are preferentially independent of X_3 iff
  preference between $\langle x_1, x_2, x_3 \rangle$ and $\langle x_1', x_2', x_3 \rangle$ does not depend on x_3
  <20,000 in flight path, $4.6 billion cost, 0.06 deaths/mpm>
  <70,000 in flight path, $4.6 billion cost, 0.06 deaths/mpm>
- If every pair of attributes is preferentially independent of its complement, then every subset of attributes is PI of its complement: mutual preferential independence
- Mutual PI $\Rightarrow \exists$ some additive value function, and we can assess n single-attribute functions as a good approximation

So what are the Complications?

- Outcome space is large
  - Like all of our problems, state spaces can be large
  - Don't want to spell out distributions like $\Pr_d$ explicitly
  - Solution: Bayes nets (or related: influence diagrams)
- Decision space is HUGE
  - Usually our decisions are not one-shot actions
  - Rather they involve sequential choices (like plans)
  - If we treat each plan as a distinct decision, decision space is too large to handle directly
  - Solution: Use dynamic programming methods to construct optimal plans (actually generalizations of plans, called policies … like in game trees)

Value of Information

- Idea: compute value of acquiring each piece of evidence
  Can be done directly from decision network
- Example: Buying oil drilling rights
  Two plots A and B only one has oil, work k
  Prior probabilities 0.50 each, mutually exclusive
  Current price of each block is k/2
  "Consultant" offers accurate survey of A. What is a fair price?
- Solution: compute expected value of information = exected value of best action given the information – expected value of best action without information
- Survey must say "oil in A" or "not oil in A"
  = [0.5 x value of "buy A" given "oil in A" + 0.5 x value of "buy B" given "not oil in A"] – 0
  = (0.5 x k/2) + (0.5 x k/2) – 0 = k/2

General Formula

- Current evidence E, current best action a
  Possible action outcome S_i, potential new evidence E_j

$$EU(\alpha \mid E) = \max_a \sum_i U(S_i) \Pr(S_i \mid E, a)$$

- Suppose we knew $E_j = e_{jk}$, then we would choose $\alpha_{e_{jk}}$ s.t.

$$EU(\alpha_{e_{jk}} \mid E, E_j = e_{jk}) = \max_a \sum_i U(S_i) \Pr(S_i \mid E, a, E_j = e_{jk})$$

- E_j is a random variable whose value is currently unknown
  $\Rightarrow$ must compute expected gain over all possible values:

$$EU(\alpha_{e_{jk}} \mid E, E_j = e_{jk}) = \max_a \sum_i U(S_i) \Pr(S_i \mid E, a, E_j = e_{jk})$$

(VPI = Value of Perfect Information)