5 Stochastic Methods

Note: The slides include Section 9.3

See the last slide for additional references for the slides

Probability Theory

The nonmonotonic logics we covered introduce a mechanism for the systems to believe in propositions (jump to conclusions) in the face of uncertainty. When the truth value of a proposition p is unknown, the system can assign one to it based on the rules in the KB.

Probability theory takes this notion further by allowing graded beliefs. In addition, it provides a theory to assign beliefs to relations between propositions (e.g., $p \wedge q$), and related propositions (the notion of dependency).

Probabilities for propositions

We write ***probability(A)***,or frequently ***P(A)*** in short, to mean the "probability of A."

But what does P(A) mean?

P(I will draw ace of hearts)

P(the coin will come up heads)

P(it will snow tomorrow)

P(the sun will rise tomorrow)

P(the problem is in the third cylinder)

P(the patient has measles)

Frequency interpretation

- Draw a card from a regular deck: 13 hearts, 13 spades, 13 diamonds, 13 clubs.
Total number of cards = n = 52 = h + s + d + c.

- The probability that the proposition
A="the card is a hearts"
is true corresponds to the relative frequency with which we expect to draw a hearts.

$$P(A) = h / n$$

Frequency interpretation (cont'd)

- **The probability of an event A is the occurrences where A holds divided by all the possible occurrences:**

P(A) = #A holds / #total

- **P (I will draw ace of hearts) ?**
- **P (I will draw a spades) ?**
- **P (I will draw a hearts or a spades) ?**
- **P (I will draw a hearts and a spades) ?**

Definitions

- **An *elementary event* or *atomic event* is a happening or occurrence that cannot be made up of other events.**
- **An *event* is a set of elementary events.**
- **The set of all possible outcomes of an event E is the *sample space* or *universe* for that event.**
- **The *probability of an event E* in a sample space S is the ratio of the number of elements in E to the total number of possible outcomes of the sample space S of E.**
Thus, P(E) = |E| / |S|.

Subjective interpretation

- **There are many situations in which there is no objective frequency interpretation:**
  - **On a cold day, just before letting myself glide from the top of Mont Ripley, I say "there is probability 0.2 that I am going to have a broken leg".**
  - **You are working hard on your AI class and you believe that the probability that you will get an A is 0.9.**
- **The probability that proposition A is true corresponds to the degree of subjective belief.**

Axioms of probability

- **There is a debate about which interpretation to adopt. But there is general agreement about the underlying mathematics.**
- **Values for probabilities should satisfy the three basic requirements:**
  - $0 \leq P(A) \leq 1$
  - $P(A \vee B) = P(A) + P(B)$
  - $P(true) = 1$

Probabilities must lie between 0 and 1

- Every probability P(A) must be positive, and between 0 and 1, inclusive: $0 \leq P(A) \leq 1$
- In informal terms it simply means that nothing can have more than a 100% chance of occurring or less than a 0% chance

Probabilities must add up

- Suppose two events are ***mutually exclusive*** i.e., only one can happen, not both
- The probability that one or the other occurs is then the sum of the individual probabilities
- Mathematically, if A and B are disjoint, i.e., $\neg (A \wedge B)$ then: $P(A \vee B) = P(A) + P(B)$
- Suppose there is a 30% chance that the stock market will go up and a 45% chance that it will stay the same. It cannot do both at once, and so the probability that it will either go up or stay the same must be 75%.

Total probability must equal 1

- Suppose a set of events is mutually exclusive and collectively exhaustive. This means that one (and only one) of the possible outcomes must occur
- The probabilities for this set of events must sum to 1
- Informally, if we have a set of events that one of them has to occur, then there is a 100% chance that one of them will indeed come to pass
- Another way of saying this is that the probability of "always true" is 1: $P(true) = 1$

These axioms are all that is needed

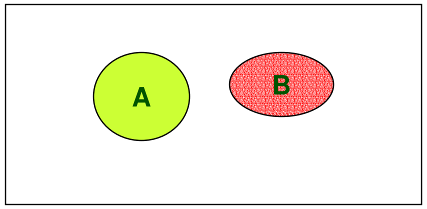
- From them, one can derive all there is to say about probabilities.
- For example we can show that:
  - $P(\neg A) = 1 - P(A)$ because

$P(A \vee \neg A) = P(true)$	by logic
$P(A \vee \neg A) = P(A) + P(\neg A)$	by the second axiom
$P(true) = 1$	by the third axiom
$P(A) + P(\neg A) = 1$	combine the above two

  - $P(false) = 0$ because

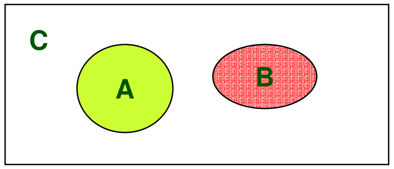
$false = \neg true$	by logic
$P(false) = 1 - P(true)$	by the above

Graphic interpretation of probability



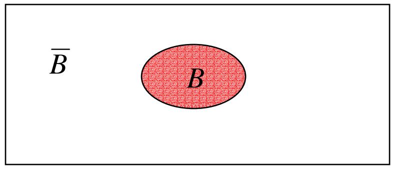
- A and B are ***events***
- They are mutually exclusive: they do not overlap, they cannot both occur at the same time
- The entire rectangle including events A and B represents everything that can occur
- Probability is represented by the area

Graphic interpretation of probability (cont'd)



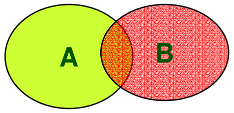
- Axiom 1: an event cannot be represented by a negative area. An event cannot be represented by an area larger than the entire rectangle
- Axiom 2: the probability of A or B occurring must be just the sum of the probability of A and the probability of B
- Axiom 3: If neither A nor B happens the event shown by the white part of the rectangle (call it C) must happen. There is a 100% chance that A, or B, or C will occur

Graphic interpretation of probability (cont'd)



- $P(\neg B) = 1 - P(B)$
- because probabilities must add to 1

Graphic interpretation of probability (cont'd)



- $P(A \lor B) = P(A) + P(B) - P(A \land B)$
- because intersection area is counted twice

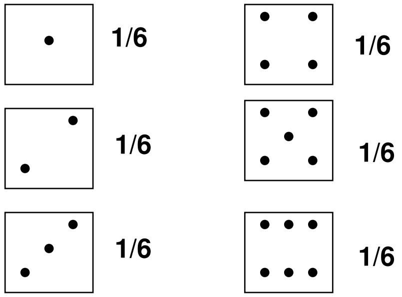
Random variables

- The events we are interested in have a set of possible values. These values are ***mutually exclusive***, and ***exhaustive***.
- For example:
  coin toss: {heads, tails}
  roll a die: {1, 2, 3, 4, 5, 6}
  weather: {snow, sunny, rain, fog}
  measles: {true, false}
- For each event, we introduce a ***random variable*** which takes on values from the associated set. Then we have:
  P(C = tails) rather than P(tails)
  P(D = 1) rather than P(1)
  P(W = sunny) rather than P(sunny)
  P(M = true) rather than P(measles)

Probability Distribution

A ***probability distribution*** is a listing of probabilities for every possible value a single random variable might take.

For example:



weather	prob.
snow	0.2
sunny	0.6
rain	0.1
fog	0.1

Joint probability distribution

A ***joint probability distribution*** for n random variables is a listing of probabilities for all possible combinations of the random variables.

For example:

Construction	Traffic	Probability
True	True	0.3
True	False	0.2
False	True	0.1
False	False	0.4

Joint probability distribution (cont'd)

Sometimes a joint probability distribution table looks like the following. It has the same information as the one on the previous slide.

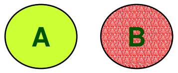
	Construction	¬Construction
Traffic	0.3	0.1
¬Traffic	0.2	0.4

Why do we need the joint probability table?

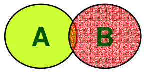
It is similar to a truth table, however, unlike in logic, it is usually not possible to derive the probability of the conjunction from the individual probabilities.

This is because the individual events interact in unknown ways. For instance, imagine that the probability of construction (C) is 0.7 in summer in Houghton, and the probability of bad traffic (T) is 0.05. If the “construction” that we are referring to in on the bridge, then a reasonable value for $P(C \wedge T)$ is 0.6. If the “construction” we are referring to is on the sidewalk of a side street, then a reasonable value for $P(C \wedge T)$ is 0.04.

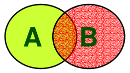
Why do we need the joint probability table? (cont’d)



$P(A \wedge B) = 0$



$P(A \wedge B) = n$



$P(A \wedge B) = m$ $m>n$

Marginal probabilities

	Construction	¬Construction	
Traffic	0.3	0.1	0.4
¬Traffic	0.2	0.4	0.6
	0.5	0.5	1.0

What is the probability of traffic, P(traffic)?

P(traffic) = P(traffic ∧ construction) + P(traffic ∧ ¬construction)
= 0.3 + 0.1
= 0.4

Note that the table should be consistent with respect to the axioms of probability: the values in the whole table should add up to 1; for any event A, P(A) should be 1 - P(¬A); and so on.

More on computing probabilities

	Construction	¬Construction	
Traffic	0.3	0.1	0.4
¬Traffic	0.2	0.4	0.6
	0.5	0.5	1.0

- Given the joint probability table, we have all the information we need about the domain. We can calculate the probability of any logical formula
- P(traffic ∨ construction) = 0.3 + 0.1 + 0.2 = 0.6
- P(construction → traffic)
  = P (¬construction ∨ traffic)   by logic
  = 0.1 + 0.4 + 0.3
  = 0.8

Dynamic probabilistic KBs

Imagine an event A. When we know nothing else, we refer to the probability of A in the usual way: P(A).

If we gather additional information, say B, the probability of A might change. This is referred to as the probability of A given B: P(A | B).

For instance, the "general" probability of bad traffic is P(T). If your friend comes over and tells you that construction has started, then the probability of bad traffic given construction is P(T | C).

Prior probability

The ***prior probability***; often called the ***unconditional probability***, of an event is the probability assigned to an event in the absence of knowledge supporting its occurrence and absence, that is, the probability of the event prior to any evidence.

The prior probability of an event is symbolized: P (event).

Posterior probability

The ***posterior*** (after the fact) probability, often called the ***conditional probability***, of an event is the probability of an event given some evidence. Posterior probability is symbolized P(event | evidence).

What are the values for the following?

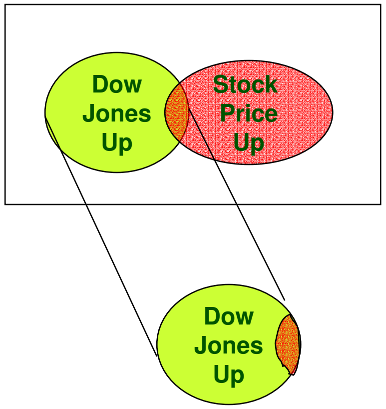
P(heads | heads)

P(ace of spades | ace of spades)

P(traffic | construction)

P(construction | traffic)

Posterior probability



Suspose that we are interested in P(up), the probability that a particular stock price will increase

Once we know that the Dow Jones has risen, then the entire rectangle is no longer appropriate

We should restrict our attention to the "Dow Jones Up" circle

Posterior probability (cont'd)

- The intuitive approach leads to the conclusion that

P (Stock Price Up given Dow Jones Up) =

P (Stock Price Up and Dow Jones Up) /

P (Dow Jones Up)

Posterior probability (cont'd)

- Mathematically, posterior probability is defined as:

$P(A \mid B) = P(A \wedge B) / P(B)$

Note that $P(B) \neq 0$.

- If we rearrange, it is called the ***product rule***:

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

Why does this make sense?

Comments on posterior probability

- P(A|B) can be thought of as:

Among all the occurrences of B, in what proportion do A and B hold together?

- If all we know is P(A), we can use this to compute the probability of A, but once we learn B, it does not make sense to use P(A) any longer.

Comparing the "conditionals"

	Construction	¬Construction	
Traffic	0.3	0.1	0.4
¬Traffic	0.2	0.4	0.6
	0.5	0.5	1.0

- P(traffic | construction)
  = $P(\text{traffic} \wedge \text{construction}) / P(\text{construction})$
  = 0.3 / 0.5 = 0.6

- $P(\text{construction} \rightarrow \text{traffic})$
  = $P(\neg\text{construction} \vee \text{traffic})$ by logic
  = 0.1 + 0.4 + 0.3
  = 0.8

- The conditional probability is usually not equal to the probability of the conditional!

Reasoning with probabilities

Pat goes in for a routine checkup and takes some tests. One test for a rare genetic disease comes back positive. The disease is potentially fatal.

She asks around and learns the following:

• rare means P(disease) = P(D) = 1/10,000

• the test is very (99%) accurate: a very small amount of false positives P(test = + | $\neg$ D) = 0.01 and no false negatives P(test = - | D) = 0.

She has to compute the probability that she has the disease and act on it. Can somebody help? Quick!!!

Making sense of the numbers

P(D) = 1/10,000

P(test = + | $\neg$ D) = 0.01 P(test = - | $\neg$ D) = 0.99

P(test = - | D) = 0, P(test = + | D) = 1

Take 10,000 people

1 will have the disease

9999 will not have the disease

1 will test positive

99.99 will test positive

9899.01 will test negative

Making sense of the numbers (cont'd)

Take 10,000 people

1 will have the disease

9999 will not have the disease

1 will test positive

99.99 will test positive
~ 100

9899.01 will test negative
~9900

P(D | test = +)

= P (D $\wedge$ test = +) / P(test = +)

= 1 / (1 + 100)

= 1 / 101 = 0.0099 ~ 0.01 (not 0.99!!)

Observe that, even if the disease were eradicated, people would test positive 1% of the time.

Formalizing the reasoning

• Bayes' rule:

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

• Apply to the example:
P(D | test= +) = P(test= + | D) P(D) / P(test= +)
= 1 * 0.0001 / P(test= +)

P($\neg$ D | test= +) = P(test= + | $\neg$ D) P($\neg$ D) / P(test= +)
= 0.01 * 0.9999 / P(test= +)

P(D | test=+) + P($\neg$D | test= +) = 1, so
P(test=+)= 0.0001 + 0.009999 = 0.010099

P (D | test= +) = 0.0001 / 0.010099 = 0.0099.

How to derive Bayes' rule

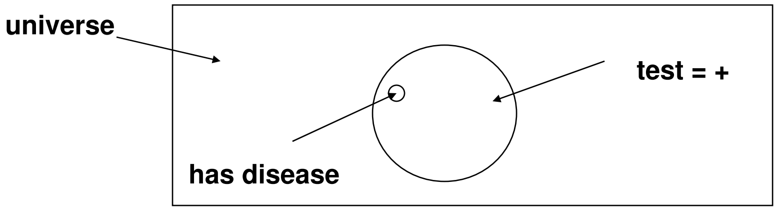
- Recall the product rule:
  $P(H \wedge E) = P(H \mid E)\, P(E)$
- $\wedge$ is commutative:
  $P(E \wedge H) = P(E \mid H)\, P(H)$
- the left hand sides are equal, so the right hand sides are too:
  $P(H \mid E)\, P(E) = P(E \mid H)\, P(H)$
- rearrange:
  $P(H \mid E) = P(E \mid H)\, P(H) / P(E)$

What did commutativity buy us?

- We can now compute probabilities that we might not have from numbers that are relatively easy to obtain.
- For instance, to compute P(measles | rash), you use P(rash|measles) and P(measles).
- Moreover, you can recompute P(measles| rash) if there is a measles epidemic and the P(measles) increases dramatically. This is more advantageous than storing the value for P(measles | rash).

What does Bayes' rule do?

It formalizes the analysis that we did for computing the probabilities:



100% of the has-disease population, i.e., those who are correctly identified as having the disease, is much smaller than 1% of the universe, i.e., those incorrectly tagged as having the disease when they don't.

Generalize to more than one evidence

- Just a piece of notation first: we use P(A, B, C) to mean $P(A \wedge B \wedge C)$.
- General form of Bayes' rule:

$$P(H \mid E_1, E_2, \dots, E_n) = P(E_1, E_2, \dots, E_n \mid H) * P(H) / P(E_1, E_2, \dots, E_n)$$

- But knowing $E_1, E_2, \dots, E_n$ requires a joint probability table for n variables. You know that this requires 2^n values.
- Can we get away with less?

Yes.

- *Independence* of some events result in simpler calculations.
Consider calculating $P(E_1, E_2, \ldots, E_n)$.
If $E_1, \ldots, E_{i-1}$ are related to weather, and $E_i, \ldots, E_n$ are related to measles, there must be some way to reason about them separately.

- Recall the coin toss example. We know that subsequent tosses are *independent*:
$P(T_1 \mid T_2) = P(T_1)$

From the product rule we have: $P(T_1 \wedge T_2) = P(T_1 \mid T_2) \times P(T_2)$.

This simplifies to $P(T_1) \times P(T_2)$ for $P(T_1 \wedge T_2)$.

Independence

- The definition of independence in terms of probability is as follows

- Events A and B are *independent* if and only if

$$P(A \mid B) = P(A)$$

- In other words, knowing whether or not B occurred will not help you find a probability for A

- For example, it seems reasonable to conclude that
P (Dow Jones Up) =
P (Dow Jones Up | It is raining in Houghton)

Independence (cont'd)

- It is important not to confuse independent events with mutually exclusive events

- Remember that two events are mutually exclusive if only one can happen at a time.

- Independent events <u>can</u> happen together

- It is possible for the Dow Jones to increase while it is raining in Houghton

Conditional independence

- This is an extension of the idea of independence

- Events A and B are said to be *conditionally independent given C*, if is it is true that
$$P(A \mid B, C) = P(A \mid C)$$

- In other words, the presence of C makes additional information B irrelevant

- If A and B are conditionally independent given C, then learning the outcome of B adds no new information regarding A if the outcome of C is already known

Conditional independence (cont'd)

- **Alternatively conditional independence means that**

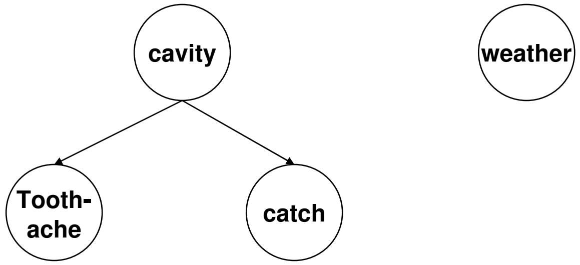
$$P(A, B \mid C) = P(A \mid C)\, P(B \mid C)$$

- **Because**

P (A , B | C)

$= P(A, B, C) / P(C)$	definition
$= P(A \mid B, C)\, P(B, C) / P(C)$	product rule
$= P(A \mid B, C)\, P(B \mid C)\, P(C) / P(C)$	product rule
$= P(A \mid B, C)\, P(B \mid C)$	cancel out P(C)
$= P(A \mid B)\, P(B \mid C)$	we had started out with assuming conditional independence

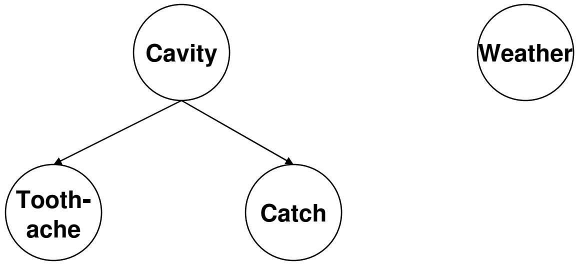
Graphically,



Cavity is the common cause of both symptoms. Toothache and cavity are independent, given a catch by a dentist with a probe:

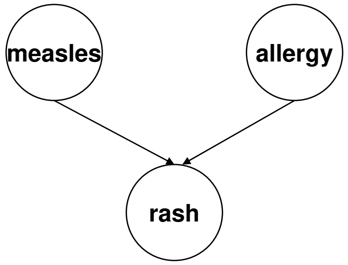
P(catch | cavity, toothache) = P(catch | cavity),
P(toothache | cavity, catch) = P(toothache | cavity).

Graphically,



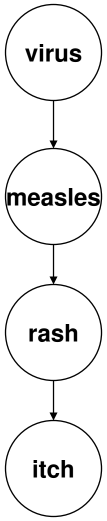
The only connection between Toothache and Catch goes through Cavity; there is no arrow directly from Toothache to Catch and vice versa

Another example



Measles and allergy influence rash independently, but if rash is given, they are dependent.

A chain of dependencies



A chain of causes is depicted here. Given measles, virus and rash are independent. In other words, once we know that the patient has measles, and evidence regarding contact with the virus is irrelevant in determining the probability of rash. Measles acts in its own way to cause the rash.

Bayesian Belief Networks (BBNs)

- **What we have just shown are *Bayesian Belief Networks* or *BBNs*. Explicitly coding the dependencies causes efficient storage and efficient reasoning with probabilities.**
- **Only probabilities of the events in terms of their parents need to be given.**
- **Some probabilities can be read off directly, some will have to be computed. Nevertheless, the full joint probability distribution table can be calculated.**
- **Next, we will define BBNs and then we will look at patterns of inference using BBNs.**

A belief network is a graph for which the following holds (Russell & Norvig, 2003)

1. A set of random variables makes up the nodes of the network. Variables may be discrete or continuous. Each node is annotated with quantitative probability information.

2. A set of directed links or arrows connects pairs of nodes. If there is an arrow from node X to node Y, X is said to be a parent of Y.

3. Each node Xi has a conditional probability distribution P(Xi | Parents (Xi)) that quantifies the effect of the parents on the node.

4. The graph has no directed cycles (and hence is a directed, acyclic graph, or DAG).

More on BBNs

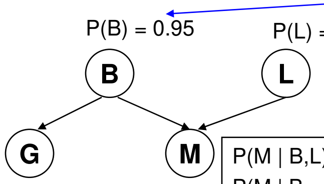
The intuitive meaning of an arrow from X to Y in a properly constructed network is usually that X has a direct influence on Y. BBNs are sometimes called *causal networks*.

It is usually easy for a domain expert to specify what direct influences exist in the domain---much easier, in fact, than actually specifying the probabilities themselves.

A Bayesian network provides a complete description of the domain.

A battery powered robot (Nilsson, 1998)

Only prior probabilities are needed for the nodes with no parents. These are the root nodes.



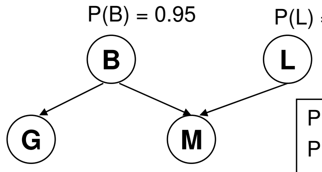
P(G|B)	= 0.95
P(G|¬B)	= 0.1

P(M | B,L)	= 0.9
P(M | B, ¬L)	= 0.05
P(M | ¬B,L)	= 0.0
P(M | ¬B, ¬ L)	= 0.0

For each leaf or intermediate node, a conditional probability table (CPT) for all the possible combinations of the parents must be given.

B: the battery is charged
L: the block is liftable
M: the robot arm moves
G: the gauge indicates that the battery is charged
(All the variables are Boolean.)

Comments on the probabilities needed



P(G|B)	= 0.95
P(G|¬B)	= 0.1

P(M | B,L)	= 0.9
P(M | B, ¬L)	= 0.05
P(M | ¬B,L)	= 0.0
P(M | ¬B, ¬ L)	= 0.0

This network has 4 variables. For the full joint probability, we would have to specify 2^4=16 probabilities (15 would be sufficient because they have to add up to 1).

In the network from, we had to specify only 8 probabilities. It does not seem like much here, but the savings are huge when n is large. The reduction can make otherwise intractable problems feasible.

Some useful rules before we proceed

- **Recall the product rule:**
  P (A ∧ B) = P(A|B) P(B)
- **We can use this to derive the *chain rule*:**

P(A, B, C, D) = P(A | B, C, D) P(B, C, D)
= P(A | B, C, D) P(B | C, D) P(C,D)
= P(A | B, C, D) P(B | C, D) P(C | D) P(D)

One can express a joint probability in terms of a chain of conditional probabilities:

P(A, B, C, D) = P(A | B, C, D) P(B | C, D) P(C | D) P(D)

Some useful rules before we proceed (cont'd)

- **How to switch variables around the conditional:**

P (A, B | C) = P(A, B, C) / P(C)

= P(A | B, C) P(B | C) P(C) / P(C) — **by the chain rule**
= P(A | B, C) P(B | C) — **delete P(C)**

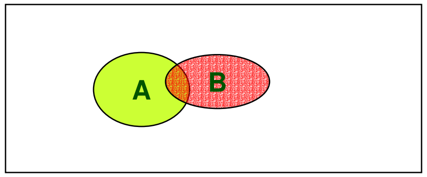
So, P (A, B | C) = P(A | B, C) P(B | C)

Total probability of an event

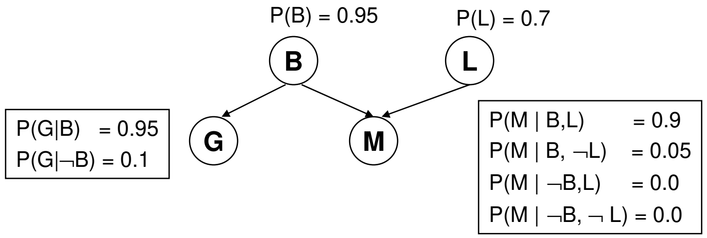
- A convenient way to calculate P(A) is with the following formula

$$P(A) = P(A \text{ and } B) + P(A \text{ and } \neg B)$$
$$= P(A \mid B)\,P(B) + P(A \mid \neg B)\,P(\neg B)$$

- Because event A is composed of those occasions when A and B occur and when A and ¬B occur. Because events “A and B” and “A and ¬B” are mutually exclusive, the probability of A must be the sum of these two probabilities



Calculating joint probabilities



What is P(G,B,M,L)?

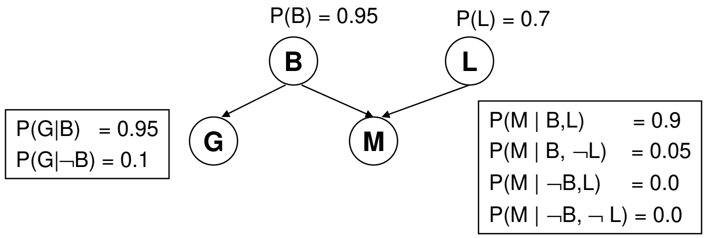
= P(G,M,B,L) — order so that lower nodes are first

= P(G|M,B,L) P(M|B,L) P(B|L) P(L) — by the chain rule

= P(G|B) P(M|B,L) P(B) P(L) — nodes need to be conditioned only on their parents

= 0.95 x 0.9 x 0.95 x 0.7 = <u>0.57</u> — read values from the BBN

Calculating joint probabilities



What is P(G,B,¬M,L)?

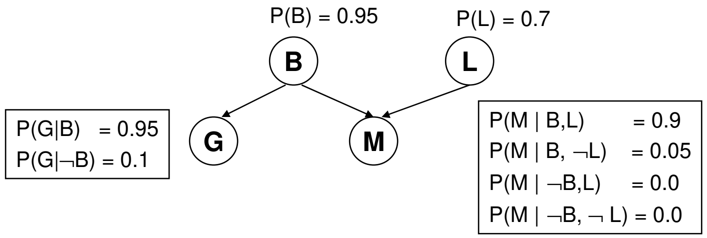
= P(G, ¬ M,B,L) — order so that lower nodes are first

= P(G| ¬ M,B,L) P(¬ M|B,L) P(B|L)P(L) — by the chain rule

= P(G|B) P(¬ M|B,L) P(B) P(L) — nodes need to be conditioned only on their parents

= 0.95 x 0.1 x 0.95 x 0.7 = <u>0.06</u> — 0.1 is 1 - 0.9

Causal or top-down inference



What is P(M | L)?

= P(M,B | L) + P(M, ¬B | L) — we want to mention the other parent too

= P(M | B,L) P(B | L) + P(M | ¬B,L) P(¬B | L) — switch around the conditional

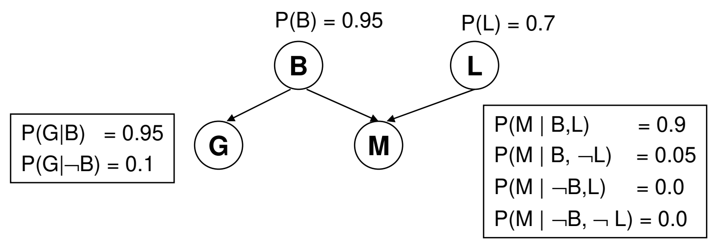
= P(M | B,L) P(B) + P(M | ¬B,L) P(¬B) — from the structure of the network

= 0.9 x 0.95 + 0 x 0.05 = <u>0.855</u>

Procedure for causal inference

- **Rewrite the desired conditional probability of the query node, V, given the evidence, in terms of the joint probability of V and all of its parents *(that are not evidence)*, given the evidence.**
- **Reexpress this joint probability back to the probability of V conditioned on all of the parents.**

Diagnostic or bottom-up inference



What is $P(\neg L \mid \neg M)$?

$= P(\neg M \mid \neg L)\, P(\neg L) / P(\neg M)$	**by Bayes' rule**
$= 0.9525 \times P(\neg L) / P(\neg M)$	**by causal inference (*)**
$= 0.9525 \times 0.3 / P(\neg M)$	**read from the table**
$= 0.9525 \times 0.3 / 0.38725 = \underline{0.7379}$	**We calculate $P(\neg M)$ by noticing that $P(\neg L \mid \neg M) + P(L \mid \neg M) = 1.0$ (***) (**)**

For (*), (**), and (***) see the following slides.

Diagnostic or bottom-up inference
(calculations needed)

- **(*) $P(\neg M \mid \neg L)$** — **use causal inference**

$= P(\neg M, B \mid \neg L) + P(\neg M, \neg B \mid L)$
$= P(\neg M \mid B, \neg L)\, P(B \mid \neg L) + P(\neg M \mid \neg B, \neg L)\, P(\neg B \mid \neg L)$
$= P(\neg M \mid B, \neg L)\, P(B) + P(\neg M \mid \neg B, \neg L)\, P(\neg B)$
$= (1 - 0.05) \times 0.95 + 1 * 0.05$
$= 0.95 * 0.95 + 0.05 = \underline{0.9525}$

- **(**) $P(L \mid \neg M)$** — **use Bayes' rule**

$= P(\neg M \mid L)\, P(L) / P(\neg M)$
$= (1 - P(M \mid L))\, P(L) / P(\neg M)$ — **$P(M|L)$ was calculated before**
$= (1 - 0.855) \times 0.7 / P(\neg M)$
$= 0.145 \times 0.7 / P(\neg M)$
$= 0.1015 / P(\neg M)$

Diagnostic or bottom-up inference
(calculations needed)

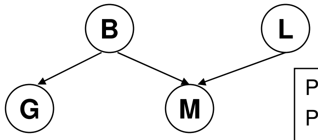
- **(***) $P(\neg L \mid \neg M) + P(L \mid \neg M) = 1$**

$0.9525 \times 0.3 / P(\neg M) + 0.145 \times 0.7 / P(\neg M) = 1$
$0.28575 / P(\neg M) + 0.1015 / P(\neg M) = 1$

$P(\neg M) = 0.38725$

Explaining away

P(B) = 0.95 P(L) = 0.7



P(G|B) = 0.95
P(G|¬B) = 0.1

P(M | B,L) = 0.9
P(M | B, ¬L) = 0.05
P(M | ¬B,L) = 0.0
P(M | ¬B, ¬ L) = 0.0

What is P(¬ L | ¬ B, ¬ M)?

= P(¬ M, ¬ B| ¬ L) P(¬ L) / P(¬ B,¬ M) — **by Bayes' rule**
= P(¬ M | ¬ B, ¬ L) P(¬ B | ¬ L) P(¬ L)/ P(¬ B,¬ M) — **switch around the conditional structure of the BBN**
= P(¬ M | ¬ B, ¬ L) P(¬ B) P(¬ L)/ P(¬ B,¬ M)
= <u>0.30</u>

Note that this is smaller than P(¬ L | ¬ M) = 0.7379 calculated before.
The additional ¬B "explained ¬L away."

Explaining away (calculations needed)

- **P(¬M | ¬B, ¬L) P(¬B | ¬L) P(¬L) / P(¬B,¬M)**
= 1 x 0.05 x 0.3 / P(¬B,¬M)
= 0.015 / P(¬B,¬M)
- **Notice that P(¬L | ¬B, ¬M) + P(L | ¬B, ¬M) must be 1.**
- **P(L | ¬B, ¬M)**
= P(¬M | ¬B, L) P(¬B | L) P(L) / P(¬B,¬M)
= 1 * 0.05 * 0.7 / P(¬B,¬M)
= 0.035 / P(¬B,¬M)
- **Solve for P(¬B,¬M).**
P(¬B,¬M) = 0.015 + 0.035 = 0.50.

Concluding remarks

- **Probability theory enables the use of varying degrees of belief to represent uncertainty**
- **A probability distribution completely describes a random variable**
- **A joint probability distribution completely describes a set of random variables**
- **Conditional probabilities let us have probabilities relative to other things that we know**
- **Bayes' rule is helpful in relating conditional probabilities and priors**

Concluding remarks (cont'd)

- **Independence assumptions let us make intractable problems tractable**
- **Belief networks are now the technology for expert systems with lots of success stories**
(e.g., Windows is shipped with a diagnostic belief network)
- **Domain experts generally report it is not to hard to interpret the links and fill in the requisite probabilities**
- **Some (e.g., Pathfinder IV) seem to be outperforming the experts consulted for their creation, some of whom are the best in the world**

Additional references used for the slides

- Jean-Claude Latombe's CS121 slides:
robotics.stanford.edu/~latombe/cs121

- Robert T. Clemen
Making Hard Decisions: An Introduction to Decision Analysis, Duxbury Press, Belmont, CA, 1990. (Chapter 7: Probability Basics)

- Nils J. Nilsson
Artificial Intelligence: A New Synthesis.
Morgan Kaufman Publishers, San Francisco, CA, 1998.

- Stuart J.Russell and Peter Norvig
Artificial Intelligence: A Modern Approach, 2nd edition.
Prentice Hall Publishers, Englewood Cliffs, NJ, 2003.