

# Uncertainty



Introduction to Artificial Intelligence

CSCI561a

Hadi Moradi

AIMA Chapter 13

(some slides are from Prof. A. Moore from CMU and Min-Yen Kan and may be updated)

# Outline

---

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

# Uncertainty

---

Let action  $A_t$  = leave for airport t minutes before flight  
Will  $A_t$  get me there on time?

Problems:

1. partial observability (road state, other drivers' plans, etc.)
2. noisy sensors (traffic reports)
3. uncertainty in action outcomes (flat tire, etc.)
4. immense complexity of modeling and predicting traffic

Hence a purely logical approach either

1. risks falsehood: " $A_{25}$  will get me there on time", or
2. leads to conclusions that are too weak for decision making:

" $A_{25}$  will get me there on time if there's no accident on the bridge and it doesn't rain and my tires remain intact etc etc."

( $A_{1440}$  might reasonably be said to get me there on time but I'd have to stay overnight in the airport ...)

# Methods for handling uncertainty

---

- Default or nonmonotonic logic:
  - Assume my car does not have a flat tire
  - Assume  $A_{25}$  works unless contradicted by evidence
- Issues: What assumptions are reasonable? How to handle contradiction?
  
- Rules with fudge factors:
  - $A_{25} \dashv\rightarrow_{0.3}$  get there on time
  - $Sprinkler \dashv\rightarrow_{0.99} WetGrass$
  - $WetGrass \dashv\rightarrow_{0.7} Rain$
- Issues: Problems with combination, e.g., *Sprinkler causes Rain??*
  
- Probability
  - Model agent's degree of belief
  - Given the available evidence,
  - $A_{25}$  will get me there on time with probability 0.04

# Probability

---

Probabilistic assertions **summarize** effects of

- **laziness**: failure to enumerate exceptions, qualifications, etc.
- **ignorance**: lack of relevant facts, initial conditions, etc.

**Subjective** probability:

- Probabilities relate propositions to agent's own state of knowledge  
e.g.,  $P(A_{25} \mid \text{no reported accidents}) = 0.06$

Probabilities of propositions change with new evidence:

e.g.,  $P(A_{25} \mid \text{no reported accidents, 5 a.m.}) = 0.15$

# Making decisions under uncertainty

---

Suppose I believe the following:

$$P(A_{25} \text{ gets me there on time} \mid \dots) = 0.04$$

$$P(A_{90} \text{ gets me there on time} \mid \dots) = 0.70$$

$$P(A_{120} \text{ gets me there on time} \mid \dots) = 0.95$$

$$P(A_{1440} \text{ gets me there on time} \mid \dots) = 0.9999$$

□ Which action to choose?

Depends on my **preferences** for missing flight vs. time spent waiting, etc.

- **Utility theory** is used to represent and infer preferences
- **Decision theory** = probability theory + utility theory

# Probabilistic Robotics

---

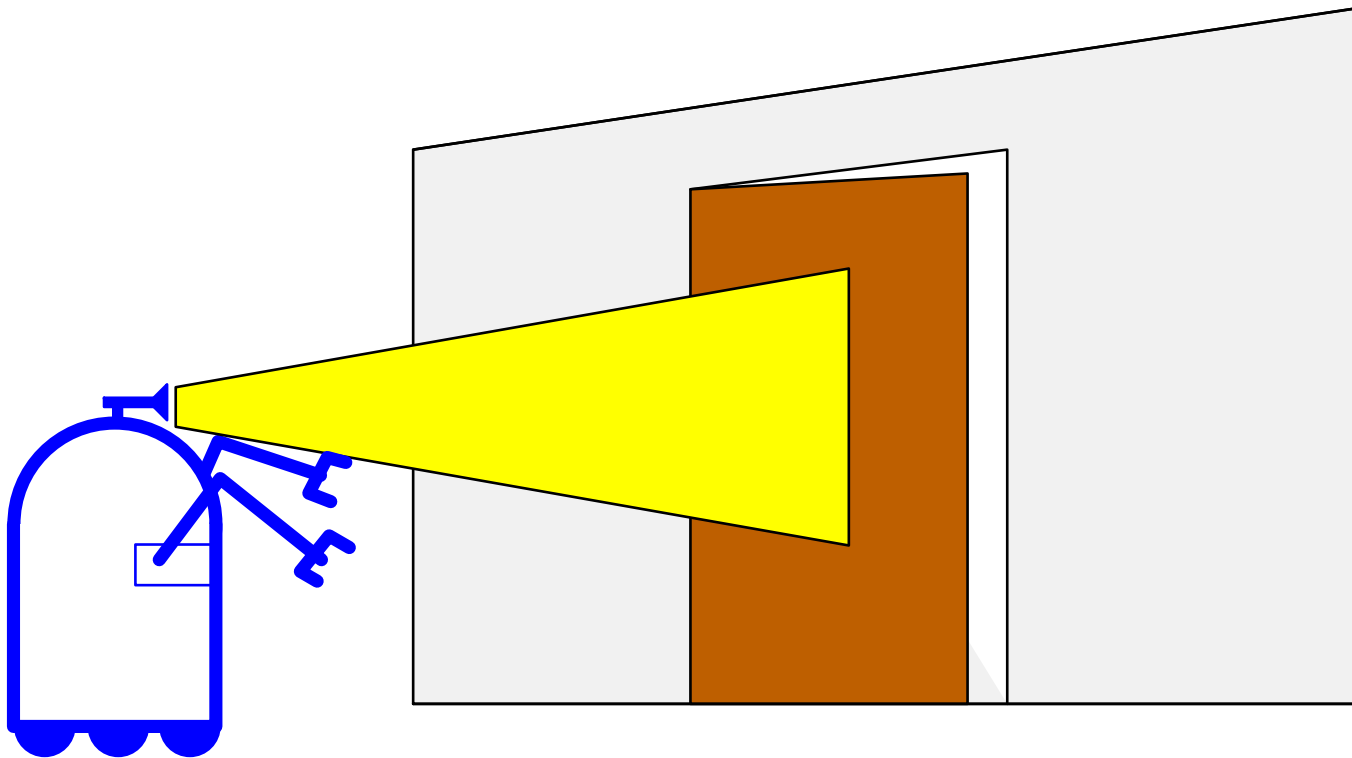
Key idea:

Explicit representation of uncertainty using the calculus of probability theory

- Perception = state estimation
- Action = utility optimization

# Another Example: State Estimation

- Suppose a robot obtains measurement  $z$
- What is  $P(open|z)$ ?



# Syntax

---

- Basic element: **random variable**
- Similar to propositional logic: possible worlds defined by assignment of values to random variables.
- **Boolean** random variables  
e.g., *Cavity* (do I have a cavity?)
- **Discrete** random variables  
e.g., *Weather* is one of  $\langle \textit{sunny}, \textit{rainy}, \textit{cloudy}, \textit{snow} \rangle$
- Domain values must be exhaustive and mutually exclusive
- Elementary proposition constructed by assignment of a value to a random variable:  
e.g.,  $\textit{Weather} = \textit{sunny}, \textit{Cavity} = \textit{false}$  (abbreviated as  $\neg \textit{cavity}$ )
- Complex propositions formed from elementary propositions and standard logical connectives e.g.,  $\textit{Weather} = \textit{sunny} \vee \textit{Cavity} = \textit{false}$

# Syntax

---

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

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

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

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

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

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

- Atomic events are mutually exclusive and exhaustive

# Probabilities

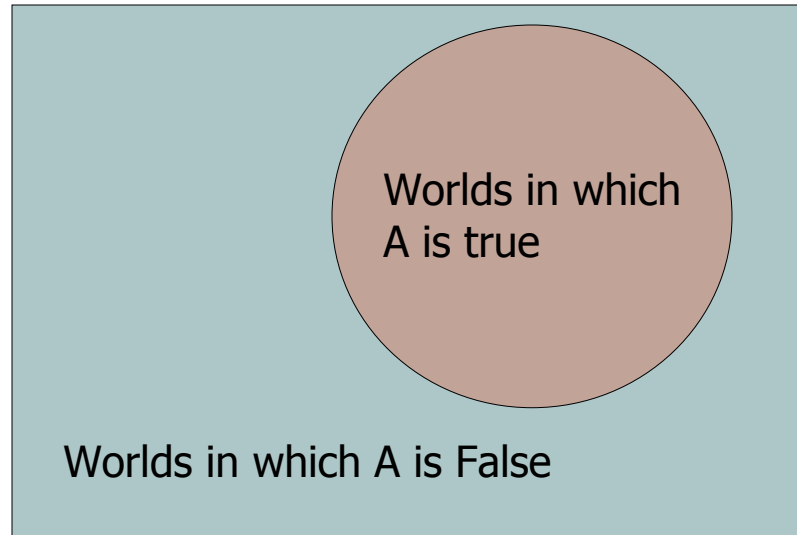
---

- We write  $P(A)$  as “the fraction of possible worlds in which  $A$  is true”

Event space of  
all possible  
worlds



Its area is 1

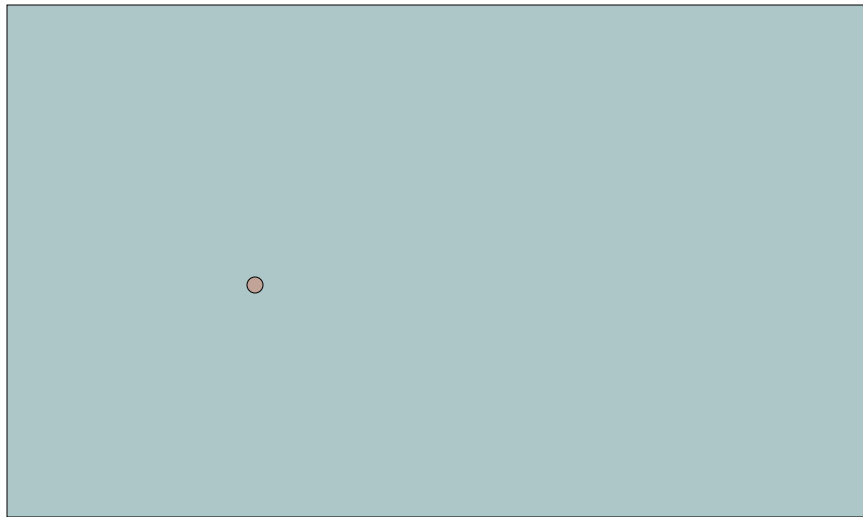


$P(A)$  = Area of  
reddish oval

# Interpreting the axioms

---

- $0 \leq P(A) \leq 1$
- $P(\text{True}) = 1$
- $P(\text{False}) = 0$



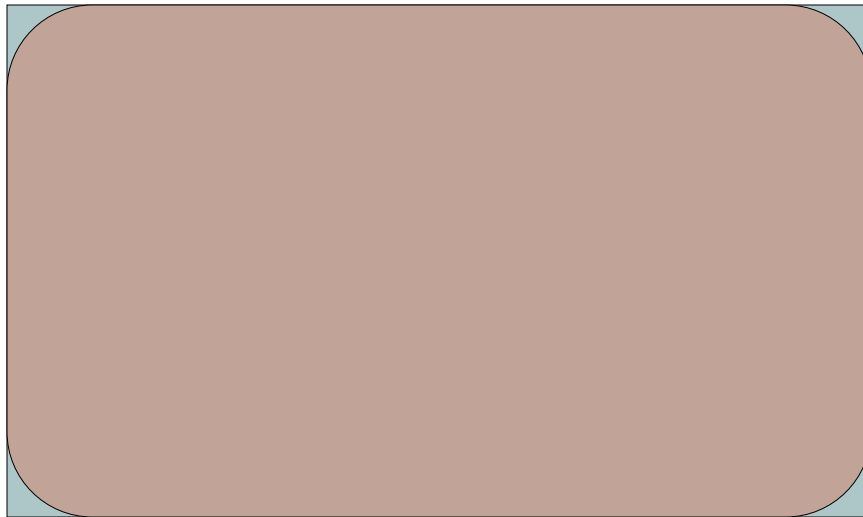
The area of A can't get any smaller than 0

And a zero area would mean no world could ever have A true

# Interpreting the axioms

---

- $0 \leq P(A) \leq 1$
- $P(\text{True}) = 1$
- $P(\text{False}) = 0$



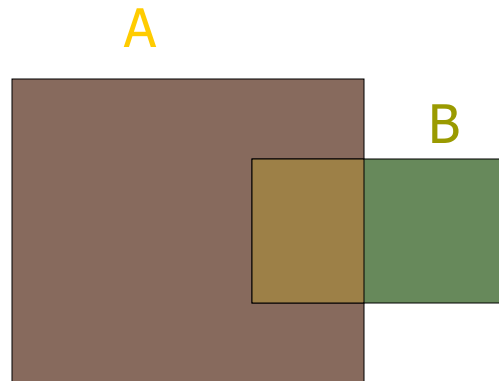
The area of A can't get any bigger than 1

And an area of 1 would mean all worlds will have A true

# Interpreting the axioms

---

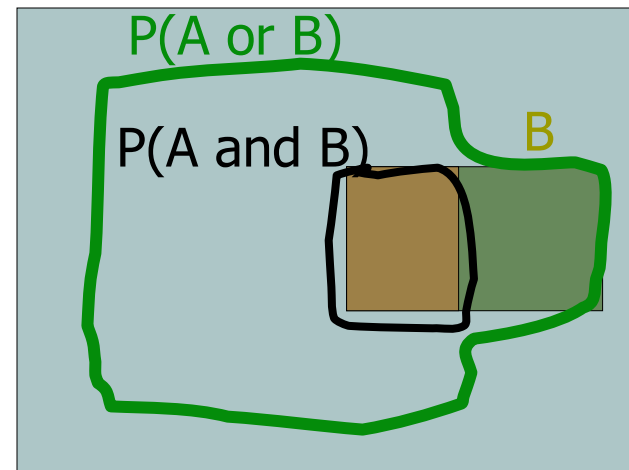
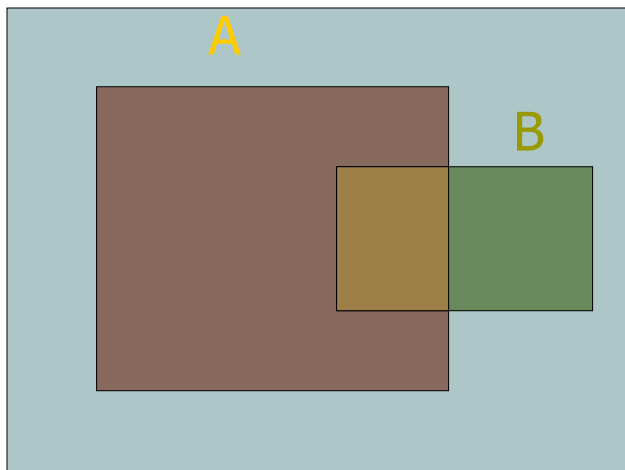
- $0 \leq P(A) \leq 1$
- $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$



# Interpreting the axioms

---

- $0 \leq P(A) \leq 1$
- $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$



Simple addition and subtraction

# Theorems from the Axioms

---

- $0 \leq P(A) \leq 1$ ,  $P(\text{True}) = 1$ ,  $P(\text{False}) = 0$
- $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

From these we can prove:

$$P(\text{not } A) = P(\sim A) = 1 - P(A)$$

- How?

# Another important theorem

---

- $0 \leq P(A) \leq 1$ ,  $P(\text{True}) = 1$ ,  $P(\text{False}) = 0$
- $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

From these we can prove:

$$P(A) = P(A \wedge B) + P(A \wedge \sim B)$$

- How?

# Multivalued Random Variables

---

- Suppose  $A$  can take on more than 2 values
- $A$  is a *random variable with arity  $k$*  if it can take on exactly one value out of  $\{v_1, v_2, \dots, v_k\}$
- Thus...

$$P(A = v_i \wedge A = v_j) = 0 \text{ if } i \neq j$$

$$P(A = v_1 \vee A = v_2 \vee \dots \vee A = v_k) = 1$$

# An easy fact about Multivalued Random Variables:

---

- Using the axioms of probability...

$$0 \leq P(A) \leq 1, P(\text{True}) = 1, P(\text{False}) = 0$$

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

- And assuming that A obeys...

$$P(A = v_i \wedge A = v_j) = 0 \text{ if } i \neq j$$

$$P(A = v_1 \vee A = v_2 \vee \dots \vee A = v_k) = 1$$

- It's easy to prove that

$$P(A = v_1 \vee A = v_2 \vee \dots \vee A = v_i) = \sum_{j=1}^i P(A = v_j)$$

# An easy fact about Multivalued Random Variables:

---

- Using the axioms of probability...

$$0 \leq P(A) \leq 1, P(\text{True}) = 1, P(\text{False}) = 0$$

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

- And assuming that A obeys...

$$P(A = v_i \wedge A = v_j) = 0 \text{ if } i \neq j$$

$$P(A = v_1 \vee A = v_2 \vee \dots \vee A = v_k) = 1$$

- It's easy to prove that

$$P(A = v_1 \vee A = v_2 \vee \dots \vee A = v_i) = \sum_{j=1}^i P(A = v_j)$$

- And thus we can prove

$$\sum_{j=1}^k P(A = v_j) = 1$$

# Another fact about Multivalued Random Variables:

---

- Using the axioms of probability...

$$0 \leq P(A) \leq 1, P(\text{True}) = 1, P(\text{False}) = 0$$

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

- And assuming that A obeys...

$$P(A = v_i \wedge A = v_j) = 0 \text{ if } i \neq j$$

$$P(A = v_1 \vee A = v_2 \vee \dots \vee A = v_k) = 1$$

- It's easy to prove that

$$P(B \wedge [A = v_1 \vee A = v_2 \vee \dots \vee A = v_i]) = \sum_{j=1}^i P(B \wedge A = v_j)$$

# Another fact about Multivalued Random Variables:

---

- Using the axioms of probability...

$$0 \leq P(A) \leq 1, P(\text{True}) = 1, P(\text{False}) = 0$$

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

- And assuming that A obeys...

$$P(A = v_i \wedge A = v_j) = 0 \text{ if } i \neq j$$

$$P(A = v_1 \vee A = v_2 \vee \dots \vee A = v_k) = 1$$

- It's easy to prove that

$$P(B \wedge [A = v_1 \vee A = v_2 \vee \dots \vee A = v_i]) = \sum_{j=1}^i P(B \wedge A = v_j)$$

- And thus we can prove

$$P(B) = \sum_{j=1}^k P(B \wedge A = v_j)$$

# Law of Total Probability, Marginals

---

## Discrete case

$$\sum_x P(x) = 1$$

$$P(x) = \sum_y P(x, y)$$

$$P(x) = \sum_y P(x | y) P(y)$$

## Continuous case

$$\int p(x) dx = 1$$

$$p(x) = \int p(x, y) dy$$

$$p(x) = \int p(x | y) p(y) dy$$

# Prior probability

---

- **Prior or unconditional probabilities** of propositions  
e.g.,  $P(\text{Cavity} = \text{true}) = 0.1$  and  $P(\text{Weather} = \text{sunny}) = 0.72$  correspond to belief prior to arrival of any (new) evidence
- **Probability distribution** gives values for all possible assignments:  
 $P(\text{Weather}) = \langle 0.72, 0.1, 0.08, 0.1 \rangle$  (**normalized**, i.e., sums to 1)
- **Joint probability distribution** for a set of random variables gives the probability of every atomic event on those random variables  
 $P(\text{Weather}, \text{Cavity}) =$  a  $4 \times 2$  matrix of values:

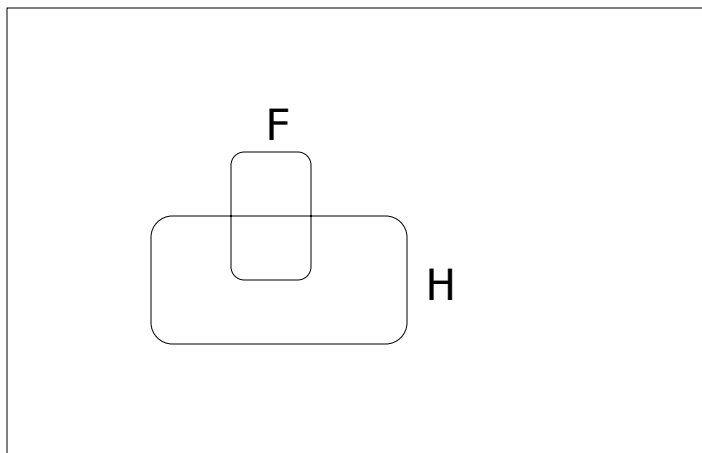
<i>Weather</i> =	sunny	rainy	cloudy	snow
<i>Cavity</i> = true	0.144	0.02	0.016	0.02
<i>Cavity</i> = false	0.576	0.08	0.064	0.08

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

# Conditional Probability

---

- $P(A|B)$  = Fraction of worlds in which B is true that also have A true



H = "Have a headache"

F = "Coming down with Flu"

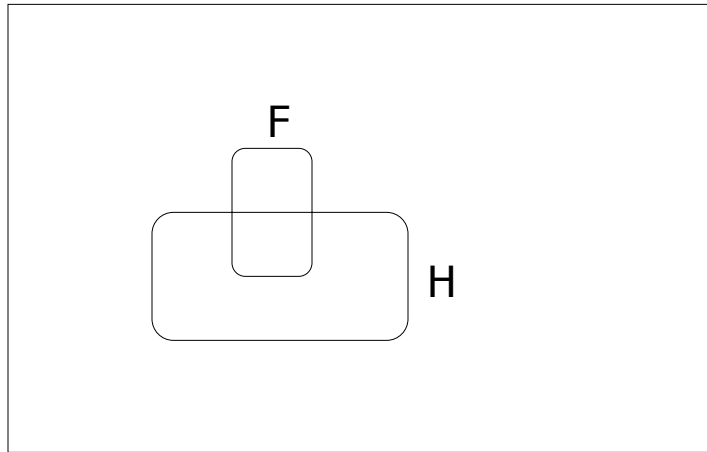
$$P(H) = 1/10$$

$$P(F) = 1/40$$

$$P(H|F) = 1/2$$

"Headaches are rare and flu is rarer, but if you're coming down with 'flu there's a 50-50 chance you'll have a headache."

# Conditional Probability



H = "Have a headache"  
F = "Coming down with Flu"

$P(H) = 1/10$   
 $P(F) = 1/40$   
 $P(H|F) = 1/2$

$P(H|F)$  = Fraction of flu-inflicted worlds in which you have a headache

$$= \frac{\text{Area of "H and F" region}}{\text{Area of "F" region}}$$

$$= \frac{P(H \wedge F)}{P(F)}$$

# Conditional probability

---

- Dentist example:
  - If we know more, e.g., *cavity* is also given, then we have
    - $P(\text{cavity} \mid \text{toothache}, \text{cavity}) = 1$
- New evidence may be irrelevant, allowing simplification, e.g.,
  - $P(\text{cavity} \mid \text{toothache}, \text{sunny}) = P(\text{cavity} \mid \text{toothache}) = 0.8$
- This kind of inference, sanctioned by domain knowledge, is crucial

# Definition of Conditional Probability

---

$$P(A/B) = \frac{P(A \wedge B)}{P(B)}$$

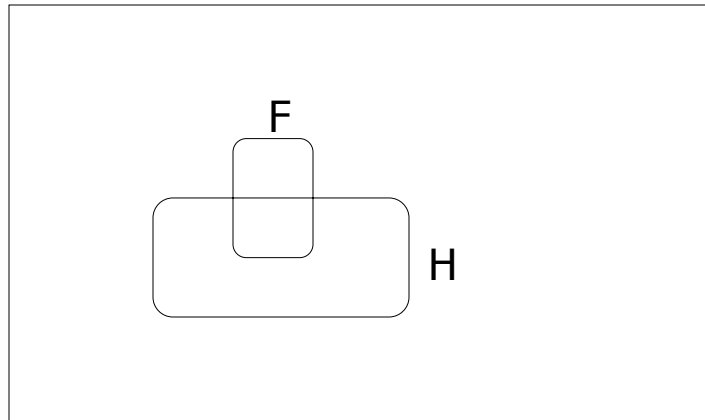
**Question:**

Any specific condition is needed here?  
What is the meaning of this condition?

## Corollary: The Chain Rule

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

# Probabilistic Inference



H = "Have a headache"

F = "Coming down with Flu"

$$P(H) = 1/10$$

$$P(F) = 1/40$$

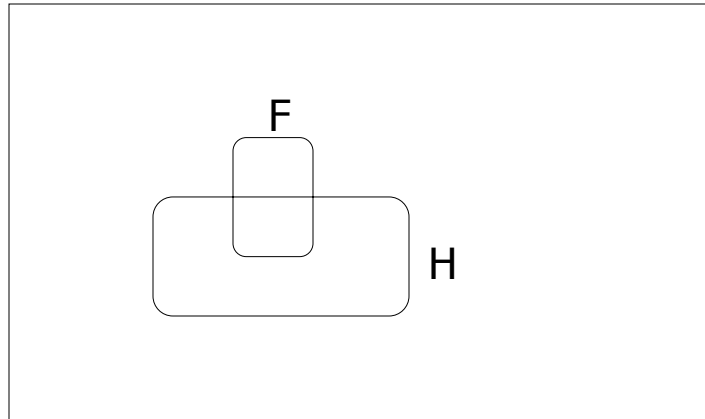
$$P(H|F) = 1/2$$

One day you wake up with a headache. You think: "Drat! 50% of flus are associated with headaches so I must have a 50-50 chance of coming down with flu"

Is this reasoning good?

# Probabilistic Inference

---



H = "Have a headache"

F = "Coming down with Flu"

$$P(H) = 1/10$$

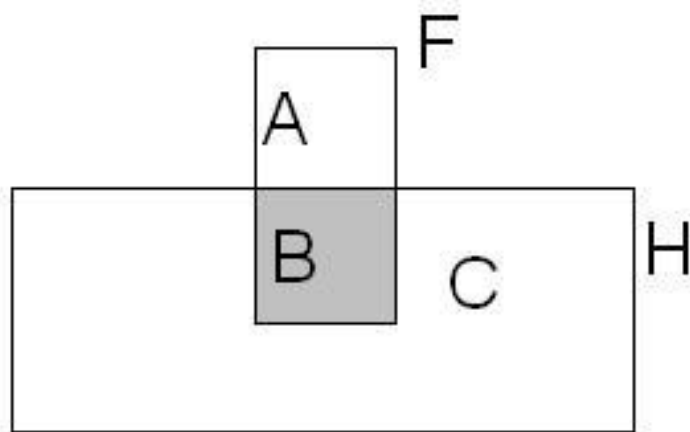
$$P(F) = 1/40$$

$$P(H|F) = 1/2$$

$$P(F \wedge H) = \dots$$

$$P(F|H) = \dots$$

# Another way to understand the intuition



Let's say we have  $P(F)$ ,  $P(H)$ , and  $P(H|F)$ , like in the example in class.

Areawise,  $P(F) = A + B$ ,  $P(H) = B + C$ ,

Also,  $P(H|F) = \frac{B}{A + B}$

Thus, to get the opposite conditional probability, ie,  $P(F|H)$ , we need to figure out  $\frac{B}{B + C}$

Since we know  $B / (A+B)$ , we can get  $B / (B+C)$  by multiplying by  $(A+B)$  and dividing by  $(B+C)$ . But since we already calculated,  $A+B = P(F)$ , and  $B+C = P(H)$ , so we are actually multiplying by  $P(F)$  and dividing by  $P(H)$ . Which is Bayes Rule:

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

Thanks to Jahanzeb Sherwani for contributing this explanation:

# What we just did...

$$P(B|A) = \frac{P(A \wedge B)}{P(A)} = \frac{P(A|B) P(B)}{P(A)}$$

$$Posterior = \frac{Prior * Likelihood}{Evidence}$$

This is Bayes Rule

**Bayes, Thomas (1763)** An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53:370-418



# Conditional probability

---

- **Product rule** gives an alternative formulation:

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

- **Chain rule** is derived by successive application of product rule:

$$\begin{aligned} P(X_1, \dots, X_n) &= P(X_1, \dots, X_{n-1}) P(X_n | X_1, \dots, X_{n-1}) \\ &= P(X_1, \dots, X_{n-2}) P(X_{n-1} | X_1, \dots, X_{n-2}) P(X_n | X_1, \dots, X_{n-1}) \\ &= \dots \\ &= \prod_{i=1}^n P(X_i | X_1, \dots, X_{i-1}) \end{aligned}$$

## More General Forms of Bayes Rule

---

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\sim A)P(\sim A)}$$

$$P(A|B \wedge X) = \frac{P(B|A \wedge X)P(A \wedge X)}{P(B \wedge X)}$$

# More General Forms of Bayes Rule

---

$$P(A = v_i | B) = \frac{P(B | A = v_i)P(A = v_i)}{\sum_{k=1}^{n_A} P(B | A = v_k)P(A = v_k)}$$

# Useful Easy-to-prove facts

---

$$P(A | B) + P(\neg A | B) = 1$$

$$\sum_{k=1}^{n_A} P(A = v_k | B) = 1$$

# Inference by enumeration

---

- Start with the joint probability distribution:

	<i>toothache</i>		$\neg$ <i>toothache</i>	
	<i>catch</i>	$\neg$ <i>catch</i>	<i>catch</i>	$\neg$ <i>catch</i>
<i>cavity</i>	<b>.108</b>	<b>.012</b>	<b>.072</b>	<b>.008</b>
$\neg$ <i>cavity</i>	<b>.016</b>	<b>.064</b>	<b>.144</b>	<b>.576</b>

- For any proposition  $\varphi$ , sum the atomic events where it is true.
- Catch: Dentist's probe caught the cavity

# Inference by enumeration

---

- Start with the joint probability distribution:

	<i>toothache</i>		$\neg$ <i>toothache</i>	
	<i>catch</i>	$\neg$ <i>catch</i>	<i>catch</i>	$\neg$ <i>catch</i>
<i>cavity</i>	<b>.108</b>	<b>.012</b>	<b>.072</b>	<b>.008</b>
$\neg$ <i>cavity</i>	<b>.016</b>	<b>.064</b>	<b>.144</b>	<b>.576</b>

- For any proposition  $\varphi$ , sum the atomic events where it is true.
- $P(\textit{toothache}) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2$

# Inference by enumeration

---

- Start with the joint probability distribution:

	<i>toothache</i>		$\neg$ <i>toothache</i>	
	<i>catch</i>	$\neg$ <i>catch</i>	<i>catch</i>	$\neg$ <i>catch</i>
<i>cavity</i>	<b>.108</b>	<b>.012</b>	<b>.072</b>	<b>.008</b>
$\neg$ <i>cavity</i>	<b>.016</b>	<b>.064</b>	<b>.144</b>	<b>.576</b>

- For any proposition  $\varphi$ , sum the atomic events where it is true.
- $P(\textit{toothache} \vee \textit{cavity}) = 0.108 + 0.012 + 0.016 + 0.064 + 0.072 + 0.008 = 0.28$

# Inference by enumeration

- Start with the joint probability distribution:

	<i>toothache</i>		$\neg$ <i>toothache</i>	
	<i>catch</i>	$\neg$ <i>catch</i>	<i>catch</i>	$\neg$ <i>catch</i>
<i>cavity</i>	<b>.108</b>	<b>.012</b>	<b>.072</b>	<b>.008</b>
$\neg$ <i>cavity</i>	<b>.016</b>	<b>.064</b>	<b>.144</b>	<b>.576</b>

- Can also compute conditional probabilities:

$$\begin{aligned} P(\neg \text{cavity} \mid \text{toothache}) &= \frac{P(\neg \text{cavity} \wedge \text{toothache})}{P(\text{toothache})} \\ &= \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064} \\ &= 0.4 \end{aligned}$$

# Normalization

	<i>toothache</i>		$\neg$ <i>toothache</i>	
	<i>catch</i>	$\neg$ <i>catch</i>	<i>catch</i>	$\neg$ <i>catch</i>
<i>cavity</i>	<b>.108</b>	<b>.012</b>	.072	.008
$\neg$ <i>cavity</i>	<b>.016</b>	<b>.064</b>	.144	.576

- Denominator can be viewed as a **normalization constant**  $\alpha$
- Probability should be limited by 1.

$$\begin{aligned} P(\text{Cavity} / \text{toothache}) &= \alpha P(\text{Cavity}, \text{toothache}) \\ &= \alpha [P(\text{Cavity}, \text{toothache}, \text{catch}) + P(\text{Cavity}, \text{toothache}, \neg \text{catch})] \\ &= \alpha [ \langle 0.108, 0.016 \rangle + \langle 0.012, 0.064 \rangle ] \\ &= \alpha \langle 0.12, 0.08 \rangle = \langle 0.6, 0.4 \rangle \end{aligned}$$

General idea: compute distribution on query variable by fixing **evidence variables** and summing over **hidden variables**

# Inference by enumeration, contd.

---

Typically, we are interested in  
the posterior joint distribution of the **query variables**  $Y$   
given specific values  $e$  for the **evidence variables**  $E$

Let the **hidden variables** be  $H = X - Y - E$

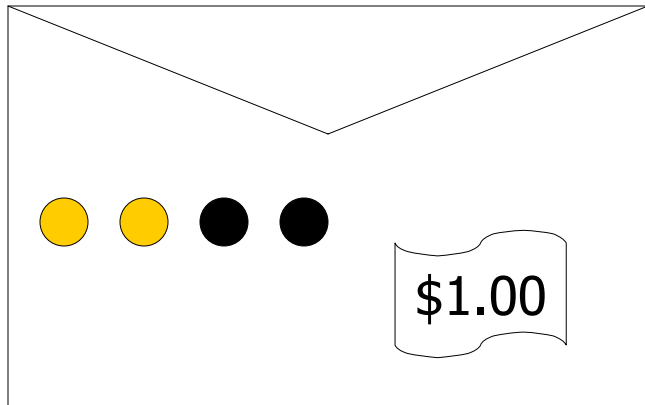
Then the required summation of joint entries is done by summing out the hidden variables:

$$P(Y | E = e) = \alpha P(Y, E = e) = \alpha \sum_{\mathbf{h}} P(Y, E = e, H = \mathbf{h})$$

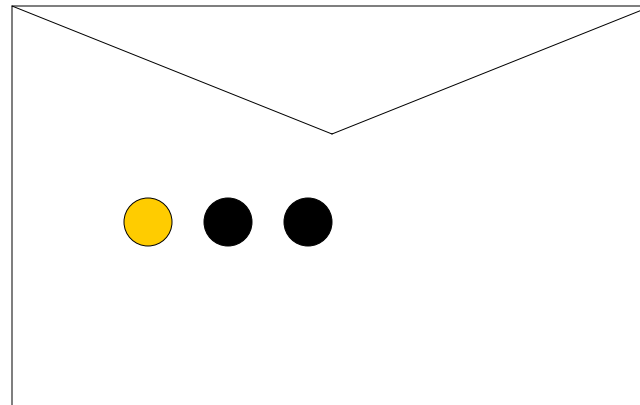
- The terms in the summation are joint entries because  $Y$ ,  $E$  and  $H$  together exhaust the set of random variables
  
- Obvious problems:
  1. Worst-case time complexity  $O(d^n)$  where  $d$  is the largest arity
  2. Space complexity  $O(d^n)$  to store the joint distribution
  3. How to find the numbers for  $O(d^n)$  entries?

# Using Bayes Rule to Gamble

---



The "Win" envelope has a dollar and four beads in it

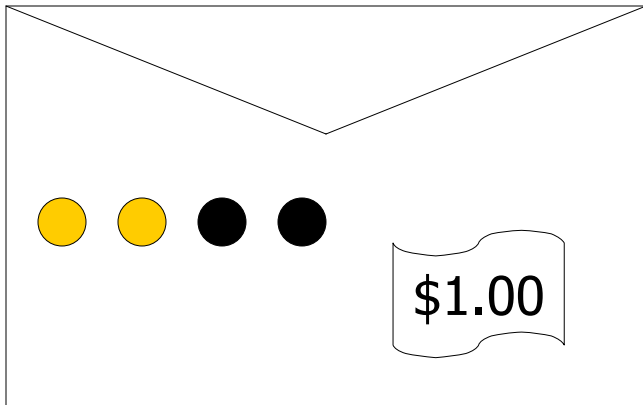


The "Lose" envelope has three beads and no money

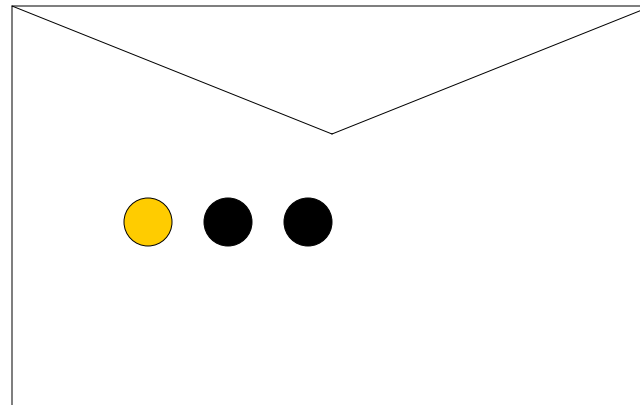
Trivial question: someone draws an envelope at random and offers to sell it to you. How much should you pay?

# Using Bayes Rule to Gamble

---



The "Win" envelope has a dollar and four beads in it



The "Lose" envelope has three beads and no money

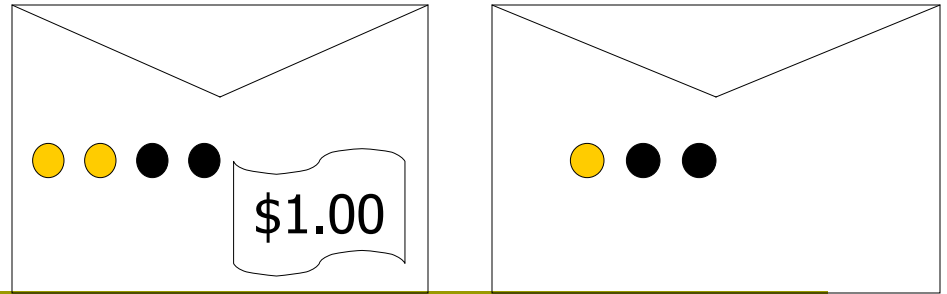
Interesting question: before deciding, you are allowed to see one bead drawn from the envelope.

Suppose it's black: How much should you pay?

Suppose it's yellow: How much should you pay?

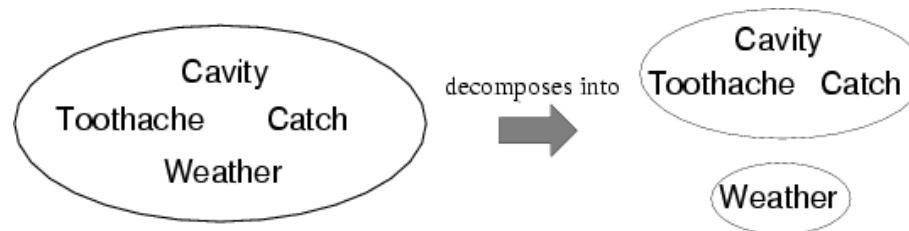
# Calculation...

---



# Independence

- $A$  and  $B$  are independent iff  
 $P(A/B) = P(A)$  or  $P(B/A) = P(B)$  or  $P(A, B) = P(A) P(B)$



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

- 32 entries reduced to 12; for  $n$  independent biased coins,  $O(2^n) \rightarrow O(n)$
- Absolute independence powerful but rare
- Dentistry is a large field with hundreds of variables, none of which are independent. What to do?

# Conditional independence

---

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

# Conditional independence contd.

---

- Write out full joint distribution using chain rule:

$$P(\textit{Toothache}, \textit{Catch}, \textit{Cavity})$$

$$= P(\textit{Toothache} \mid \textit{Catch}, \textit{Cavity}) P(\textit{Catch}, \textit{Cavity})$$

$$= P(\textit{Toothache} \mid \textit{Catch}, \textit{Cavity}) P(\textit{Catch} \mid \textit{Cavity}) P(\textit{Cavity})$$

$$= P(\textit{Toothache} \mid \textit{Cavity}) P(\textit{Catch} \mid \textit{Cavity}) P(\textit{Cavity})$$

I.e.,  $2 + 2 + 1 = 5$  independent numbers (reduction from 7 to 5)

- In most cases, the use of conditional independence reduces the size of the representation of the joint distribution from exponential in  $n$  to linear in  $n$ .
- Conditional independence is our most basic and robust form of knowledge about uncertain environments.

# Bayes' Rule

---

- Product rule  $P(a \wedge b) = P(a | b) P(b) = P(b | a) P(a)$   
⇒ **Bayes' rule**:  $P(a | b) = P(b | a) P(a) / P(b)$
- or in distribution form  
$$P(Y|X) = P(X|Y) P(Y) / P(X) = \alpha P(X|Y) P(Y)$$
- Useful for assessing **diagnostic** probability from **causal** probability:
  - $P(\text{Cause}|\text{Effect}) = P(\text{Effect}|\text{Cause}) P(\text{Cause}) / P(\text{Effect})$
  - E.g., let  $M$  be meningitis,  $S$  be stiff neck:  
 $P(m|s) = P(s|m) P(m) / P(s) = 0.8 \times 0.0001 / 0.1 = 0.0008$
  - Note: posterior probability of meningitis still very small!

# Bayes' Rule and conditional independence

$$\begin{aligned} P(\text{Cavity} \mid \text{toothache} \wedge \text{catch}) \\ &= \alpha P(\text{toothache} \wedge \text{catch} \mid \text{Cavity}) P(\text{Cavity}) \\ &= \alpha P(\text{toothache} \mid \text{Cavity}) P(\text{catch} \mid \text{Cavity}) P(\text{Cavity}) \end{aligned}$$

- This is an example of a **naïve Bayes** model:

$$P(\text{Cause}, \text{Effect}_1, \dots, \text{Effect}_n) = P(\text{Cause}) \prod_i P(\text{Effect}_i \mid \text{Cause})$$



- Total number of parameters is **linear** in  $n$

# The Joint Distribution

*Example: Boolean  
variables A, B, C*

---

Recipe for making a joint distribution  
of M variables:

# The Joint Distribution

*Example: Boolean variables A, B, C*

Recipe for making a joint distribution of M variables:

1. Make a truth table listing all combinations of values of your variables (if there are M Boolean variables then the table will have  $2^M$  rows).

A	B	C
0	0	0
0	0	1
0	1	0
0	1	1
1	0	0
1	0	1
1	1	0
1	1	1

# The Joint Distribution

*Example: Boolean variables A, B, C*

Recipe for making a joint distribution of M variables:

1. Make a truth table listing all combinations of values of your variables (if there are M Boolean variables then the table will have  $2^M$  rows).
2. For each combination of values, say how probable it is.

A	B	C	Prob
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10

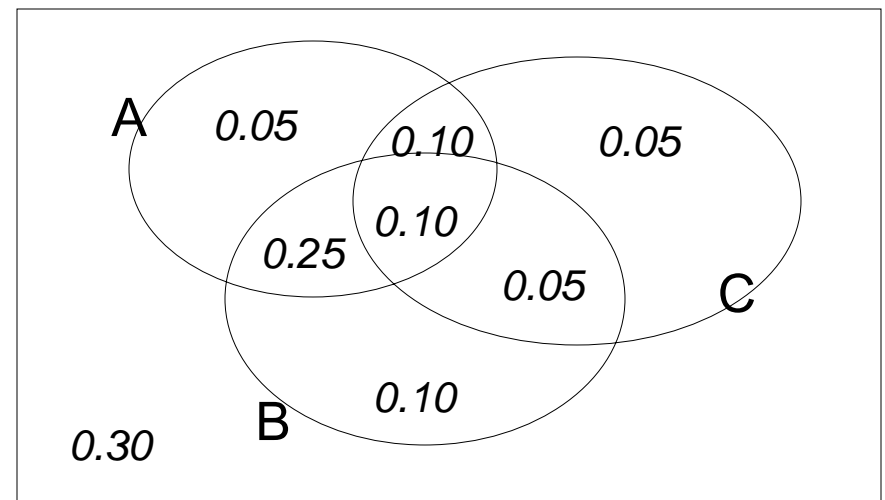
# The Joint Distribution

*Example: Boolean variables A, B, C*









Recipe for making a joint distribution of M variables:

1. Make a truth table listing all combinations of values of your variables (if there are M Boolean variables then the table will have  $2^M$  rows).
2. For each combination of values, say how probable it is.
3. If you subscribe to the axioms of probability, those numbers must sum to 1.

A	B	C	Prob
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10









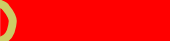

# Using the Joint

gender	hours_worked	wealth		
Female	v0:40.5-	poor	0.253122	
		rich	0.0245895	
	v1:40.5+	poor	0.0421768	
		rich	0.0116293	
Male	v0:40.5-	poor	0.331313	
		rich	0.0971295	
	v1:40.5+	poor	0.134106	
		rich	0.105933	

Once you have the JD you can ask for the probability of any logical expression involving your attribute

$$P(E) = \sum_{\text{rows matching } E} P(\text{row})$$

# Using the Joint

gender	hours_worked	wealth		
Female	v0:40.5-	poor	0.253122	
		rich	0.0245895	
	v1:40.5+	poor	0.0421768	
		rich	0.0116293	
Male	v0:40.5-	poor	0.331313	
		rich	0.0971295	
	v1:40.5+	poor	0.134106	
		rich	0.105933	

$$P(\text{Poor Male}) = 0.4654$$

$$P(E) = \sum_{\text{rows matching } E} P(\text{row})$$









# Using the Joint

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

$$P(\text{Poor}) = 0.7604$$

$$P(E) = \sum_{\text{rows matching } E} P(\text{row})$$

# Inference with the Joint

gender	hours_worked	wealth		
Female	v0:40.5-	poor	0.253122	
		rich	0.0245895	
	v1:40.5+	poor	0.0421768	
		rich	0.0116293	
Male	v0:40.5-	poor	0.331313	
		rich	0.0971295	
	v1:40.5+	poor	0.134106	
		rich	0.105933	

$$P(E_1 | E_2) = \frac{P(E_1 \wedge E_2)}{P(E_2)} = \frac{\sum_{\text{rows matching } E_1 \text{ and } E_2} P(\text{row})}{\sum_{\text{rows matching } E_2} P(\text{row})}$$

# Inference with the Joint

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

$$P(E_1 | E_2) = \frac{P(E_1 \wedge E_2)}{P(E_2)} = \frac{\sum_{\text{rows matching } E_1 \text{ and } E_2} P(\text{row})}{\sum_{\text{rows matching } E_2} P(\text{row})}$$

$$P(\text{Male} | \text{Poor}) = 0.4654 / 0.7604 = 0.612$$

# Joint distributions

---

- Good news

Once you have a joint distribution, you can ask important questions about stuff that involves a lot of uncertainty

- Bad news

Impossible to create for more than about ten attributes because there are so many numbers needed when you build the damn thing.

# Using fewer numbers

---

Suppose there are two events:

- M: Manuela teaches the class (otherwise it's Hadi)
- S: It is sunny

The joint p.d.f. for these events contain four entries.

If we want to build the joint p.d.f. we'll have to invent those four numbers.

- Having  $P(M)$  and  $P(S)$  don't derive the joint distribution. So you can't answer all questions.

What extra assumption can you make?

# Independence

---

“The sunshine levels do not depend on and do not influence who is teaching.”

This can be specified very simply:

$$P(S \mid M) = P(S)$$

This is a powerful statement!

It required extra domain knowledge. A different kind of knowledge than numerical probabilities. It needed an understanding of causation.

# Independence

---

From  $P(S \mid M) = P(S)$ , the rules of probability imply: (*can you prove these?*)

1.  $P(\sim S \mid M) = P(\sim S)$
2.  $P(M \mid S) = P(M)$
3.  $P(M \wedge S) = P(M) P(S)$
4.  $P(\sim M \wedge S) = P(\sim M) P(S)$
5.  $P(M \wedge \sim S) = P(M) P(\sim S)$
6.  $P(\sim M \wedge \sim S) = P(\sim M) P(\sim S)$

# Independence

From  $P(S | M) = P(S)$ , the rules of probability imply: (*can you prove these?*)

And in general:

- $P(M=u \wedge S=v) = P(M=u) P(S=v)$
- $P(M)$  for each of the four combinations of
  - $u=True/False$
  - $v=True/False$
- $P(\sim M \wedge \sim S) = P(\sim M)P(\sim S)$

# Independence

---

We've stated:

$$P(M) = 0.6$$

$$P(S) = 0.3$$

$$P(S \mid M) = P(S)$$

From these statements, we can derive the full joint pdf.

M	S	Prob
T	T	
T	F	
F	T	
F	F	

And since we now have the joint pdf, we can make any queries we like.

# A more interesting case

---

- M : Manuela teaches the class
- S : It is sunny
- L : The lecturer arrives slightly late.

Assume both lecturers are sometimes delayed by bad weather. Hadi is more likely to arrive later than Manuela.

# A more interesting case

---

- M : Manuela teaches the class
- S : It is sunny
- L : The lecturer arrives slightly late.

Assume both lecturers are sometimes delayed by bad weather. Hadi is more likely to arrive later than Manuela.

Let's begin with writing down knowledge we're happy about:

$$P(S \mid M) = P(S), \quad P(S) = 0.3, \quad P(M) = 0.6$$

Lateness is not independent of the weather and is not independent of the lecturer.

# A more interesting case

---

- M : Manuela teaches the class
- S : It is sunny
- L : The lecturer arrives slightly late.

Assume both lecturers are sometimes delayed by bad weather. Hadi is more likely to arrive later than Manuela.

Let's begin with writing down knowledge we're happy about:

$P(S \mid M) = P(S)$ ,  $P(S) = 0.3$ ,  $P(M) = 0.6$   
Lateness is not independent of the weather and is not independent of the lecturer.

We already know the Joint of S and M, so all we need now is

$$P(L \mid S=u, M=v)$$

in the 4 cases of  $u/v = \text{True/False}$ .

# A more interesting case

---

- M : Manuela teaches the class
- S : It is sunny
- L : The lecturer arrives slightly late.

Assume both lecturers are sometimes delayed by bad weather. Hadi is more likely to arrive later than Manuela.

$$\begin{array}{l} P(S \mid M) = P(S) \\ P(S) = 0.3 \\ P(M) = 0.6 \end{array} \quad \begin{array}{l} P(L \mid M \wedge S) = 0.05 \\ P(L \mid M \wedge \sim S) = 0.1 \\ P(L \mid \sim M \wedge S) = 0.1 \\ P(L \mid \sim M \wedge \sim S) = 0.2 \end{array}$$

Now we can derive a full joint p.d.f. with a “mere” six numbers instead of seven\*

*\*Savings are larger for larger numbers of variables.*

# A more interesting case

- M : Manuela teaches the class
- S : It is sunny
- L : The lecturer arrives slightly late.

Assume both lecturers are sometimes delayed by bad weather. Hadi is more likely to arrive later than Manuela.

$$P(S \mid M) = P(S)$$

$$P(S) = 0.3$$

$$P(M) = 0.6$$

$$P(L \mid M \wedge S) = 0.05$$

$$P(L \mid M \wedge \sim S) = 0.1$$

$$P(L \mid \sim M \wedge S) = 0.1$$

$$P(L \mid \sim M \wedge \sim S) = 0.2$$

Question: Express

$$P(L=x \wedge M=y \wedge S=z)$$

in terms that only need the above expressions, where  $x, y$  and  $z$  may each be True or False.

# A bit of notation

$$P(S | M) = P(S)$$

$$P(S) = 0.3$$

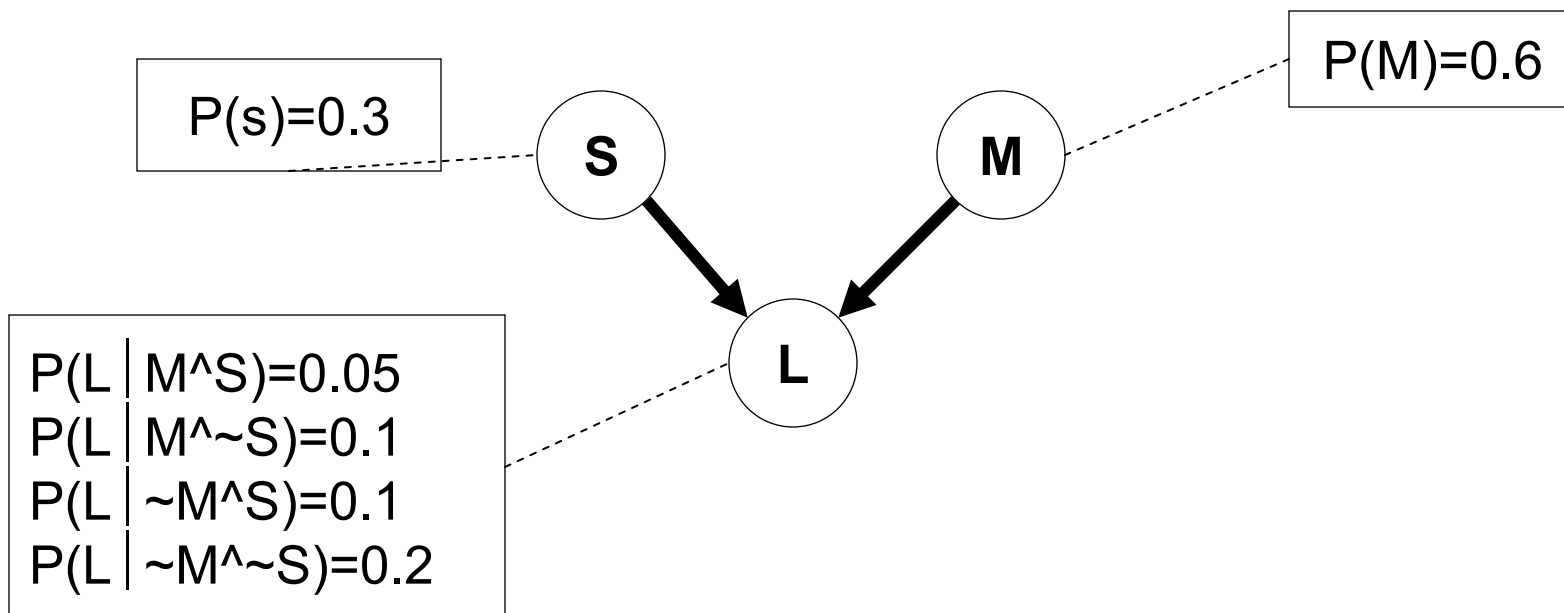
$$P(M) = 0.6$$

$$P(L | M \wedge S) = 0.05$$

$$P(L | M \wedge \sim S) = 0.1$$

$$P(L | \sim M \wedge S) = 0.1$$

$$P(L | \sim M \wedge \sim S) = 0.2$$



# A bit of notation

This kind of stuff will be thoroughly formalized later

$$P(S | M) = P(S)$$

$$P(S) = 0.3$$

$$P(M) = 0.6$$

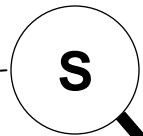
$$P(L | M)$$

$$P(L | \sim M)$$

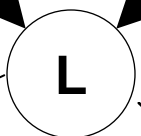
$$P(L | \sim M)$$

Read the absence of an arrow between S and M to mean "it would not help me predict M if I knew the value of S"

$$P(s)=0.3$$



$$P(M)=0.6$$



$$P(L | M \wedge S) = 0.05$$

$$P(L | M \wedge \sim S) = 0.1$$

$$P(L | \sim M \wedge S) = 0.1$$

$$P(L | \sim M \wedge \sim S) = 0.2$$

Read the two arrows into L to mean that if I want to know the value of L it may help me to know M and to know S.

# An even cuter trick

---

Suppose we have these three events:

- M : Lecture taught by Manuela
- L : Lecturer arrives late
- R : Lecture concerns robots

Suppose:

- Hadi has a higher chance of being late than Manuela.
- Hadi has a higher chance of giving robotics lectures.

What kind of independence can we find?

How about:

- $P(L \mid M) = P(L) ?$
- $P(R \mid M) = P(R) ?$
- $P(L \mid R) = P(L) ?$

# Conditional independence

---

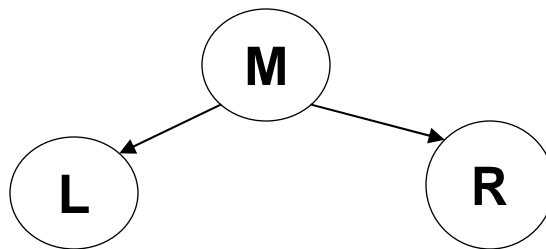
Once you know who the lecturer is, then whether they arrive late doesn't affect whether the lecture concerns robots.

$$P(R \mid M, L) = P(R \mid M) \text{ and}$$
$$P(R \mid \sim M, L) = P(R \mid \sim M)$$

We express this in the following way:

“R and L are conditionally independent given M”

..which is also notated by the following diagram.



Given knowledge of M, knowing anything else in the diagram won't help us with L, etc.

# Conditional Independence formalized

---

R and L are conditionally independent given M if for all  $x, y, z$  in  $\{T, F\}$ :

$$P(R=x \mid M=y \wedge L=z) = P(R=x \mid M=y)$$

More generally:

Let  $S_1$  and  $S_2$  and  $S_3$  be sets of variables.

Set-of-variables  $S_1$  and set-of-variables  $S_2$  are **conditionally independent given  $S_3$**  if for all assignments of values to the variables in the sets,

$$P(S_1\text{'s assignments} \mid S_2\text{'s assignments} \ \& \ S_3\text{'s assignments}) = P(S_1\text{'s assignments} \mid S_3\text{'s assignments})$$

## Example:

“Shoe-size is conditionally independent of Glove-size given height weight and age”

means

forall  $s, g, h, w, a$

$$P(\text{ShoeSize}=s | \text{Height}=h, \text{Weight}=w, \text{Age}=a)$$

=

$$P(\text{ShoeSize}=s | \text{Height}=h, \text{Weight}=w, \text{Age}=a, \text{GloveSize}=g)$$

R and L are  
for all  $x, y, z$

$P(R)$

More gener

Let  $S_1$  and  $S_2$  and  $S_3$  be sets of variable

Set-of-variables  $S_1$  and set-of-variables  $S_2$  are **conditionally independent given  $S_3$**  if for all assignments of values to the variables in the sets,

$$P(S_1\text{'s assignments} \mid S_2\text{'s assignments} \ \& \ S_3\text{'s assignments}) = P(S_1\text{'s assignments} \mid S_3\text{'s assignments})$$

## Example:

“Shoe-size is conditionally independent of Glove-size given height weight and age”

does not mean

forall s,g,h

$$P(\text{ShoeSize}=s|\text{Height}=h)$$

=

$$P(\text{ShoeSize}=s|\text{Height}=h, \text{GloveSize}=g)$$

R and L are  
for all x,y,z

P(R

More gener

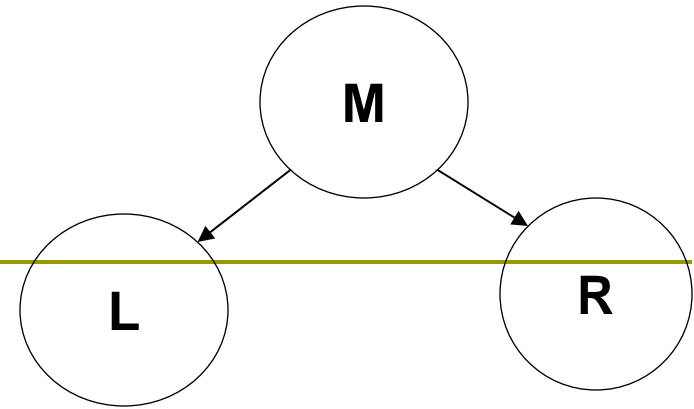
Let S1 and S2 and S3 be sets of variable

Set-of-variables S1 and set-of-variables S2 are **conditionally independent given S3** if for all assignments of values to the variables in the sets,

$$P(S_1\text{'s assignments} \mid S_2\text{'s assignments} \ \& \ S_3\text{'s assignments}) = P(S_1\text{'s assignments} \mid S_3\text{'s assignments})$$

# Conditional independence

---



We can write down  $P(M)$ . And then, since we know L is only directly influenced by M, we can write down the values of  $P(L | M)$  and  $P(L | \sim M)$  and know we've fully specified L's behavior. Ditto for R.

$$P(M) = 0.6$$

$$P(L | M) = 0.085$$

$$P(L | \sim M) = 0.17$$

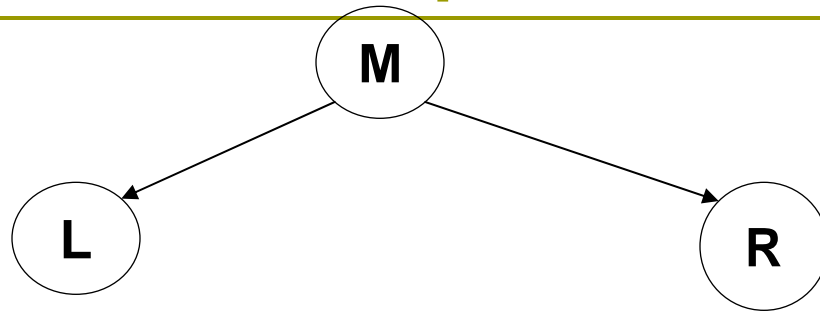
$$P(R | M) = 0.3$$

$$P(R | \sim M) = 0.6$$

'R and L conditionally independent given M'

# Conditional independence

---



$$P(M) = 0.6$$

$$P(L \mid M) = 0.085$$

$$P(L \mid \sim M) = 0.17$$

$$P(R \mid M) = 0.3$$

$$P(R \mid \sim M) = 0.6$$

Conditional Independence:

$$P(R \mid M, L) = P(R \mid M),$$

$$P(R \mid \sim M, L) = P(R \mid \sim M)$$

Again, we can obtain any member of the Joint prob dist that we desire:

$$P(L=x \wedge R=y \wedge M=z) =$$

# Summary

---

- Probability is a rigorous formalism for uncertain knowledge
- **Joint probability distribution** specifies probability of every **atomic event**
- Queries can be answered by summing over atomic events
- For nontrivial domains, we must find a way to reduce the joint size
  - Use **Independence** and **conditional independence** provide the tools