# CS:4420 Artificial Intelligence

## Spring 2019

# Uncertainty

Cesare Tinelli

The University of Iowa

# Readings

- Chap. 13 of [Russell and Norvig, 3rd Edition]

# Logic and Uncertainty

Major problem with logical-agent approaches:

> *Agents almost never have access to the whole truth about their environments*

- Very often, even in simple worlds, there are important questions for which there is no yes/no answer
- In that case, an agent must reason under uncertainty
- Uncertainty also arises because of an agent's incomplete or incorrect understanding of its environment

# Uncertainty

Let action $A_t$ = "leave for airport $t$ minutes before flight"

Will $A_t$ get me there on time?

Problems

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

# Uncertainty

Let action $A_t$ = "leave for airport $t$ minutes before flight"

Will $A_t$ get me there on time?

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 way, it doesn't rain, my tires remain intact, . . .")

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

# Reasoning under Uncertainty

A *rational* agent is one that makes rational decisions — in order to maximize its performance measure)

A rational decision depends on

- the relative importance of various goals
- the likelihood they will be achieved
- the degree to which they will be achieved

# Handling Uncertain Knowledge

Reasons FOL-based approaches fail to cope with domains like, for instance, medical diagnosis:

- Laziness: too much work to write complete axioms, or too hard to work with the enormous sentences that result
- Theoretical Ignorance: The available knowledge of the domain is incomplete
- Practical Ignorance: The theoretical knowledge of the domain is complete but some evidential facts are missing

# Degrees of Belief

In several real-world domains the agent's knowledge can only provide a degree of belief in the relevant sentences

The agent cannot say whether a sentence is true, but only that is true $x\%$ of the times

The main tool for handling degrees of belief is *Probability Theory*

The use of *probability summarizes the uncertainty that stems from our laziness or ignorance about the domain*

# Probability Theory

Probability Theory makes the same ontological commitments as First-order Logic:

*Every sentence $\varphi$ is either true or false*

The *degree of belief* that $\varphi$ is true is a number $P$ between 0 and 1

$P(\varphi) = 1 \quad \longrightarrow \quad \varphi$ is certainly true

$P(\varphi) = 0 \quad \longrightarrow \quad \varphi$ is certainly not true

$P(\varphi) = 0.65 \quad \longrightarrow \quad \varphi$ is true with a 65% chance

# Probability of Facts

Let $A$ be a propositional variable, a symbol denoting a proposition that is either true or false

$P(A)$ denotes the probability that $A$ is true *in the absence of any other information*

Similarly,

$$
\begin{aligned}
P(\neg A) &= \text{probability that } A \text{ is false} \\
P(A \wedge B) &= \text{probability that both } A \text{ and } B \text{ are true} \\
P(A \vee B) &= \text{probability that either } A \text{ or } B \text{ (or both) are true}
\end{aligned}
$$

Examples:

$P(\neg Blonde)$ $\quad$ $P(Blonde \wedge BlueEyed)$ $\quad$ $P(Blonde \vee BlueEyed)$

where $Blonde/BlueEyed$ denotes that a given person is blonde/blue-eyed

# Subjective/Bayesian Probability

Probabilities relate propositions to one'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., $P(A_{25} \mid \text{no reported accidents, 5 a.m.}) = 0.15$

Note: This is analogous to logical entailment status $KB \models \alpha$ (which changes with more knowledge), not truth

# Conditional/Unconditional Probability

$P(A)$ is the *unconditional (or prior) probability* of fact $A$

An agent can use the unconditional probability of $A$ to reason about $A$ in the absence of further information

If further evidence $B$ becomes available, the agent must use the *conditional (or posterior) probability*:

$$P(A \mid B)$$

the probability of $A$ given that (all) the agent knows (is) $B$

**Note:** $P(A)$ can be thought as the conditional probability of $A$ with respect to the empty evidence: $P(A) = P(A \mid \,)$

# Conditional Probabilities

The probability of a fact may change as the agent acquires more, or different, information:

1. $P(Blonde)$
2. $P(Blonde \mid Swedish)$
3. $P(Blonde \mid Kenian)$
4. $P(Blonde \mid Kenian \land \neg EuroDescent)$

1. If we know nothing about a person, the probability that s/he is blonde equals a certain value, say 0.2
2. If we know that a person is Swedish the probability that s/he is blonde is much higher, say 0.9
3. If we know that the person is Kenyan, the probability s/he is blonde much lower, say 0.00003
4. If we know that the person is Kenyan and not of European descent, the probability s/he is blonde is 0

# The Axioms of Probability

Probability Theory is governed by the following axioms:

1. All probabilities are real values between 0 and 1:

   for all $\varphi$, $0 \leq P(\varphi) \leq 1$

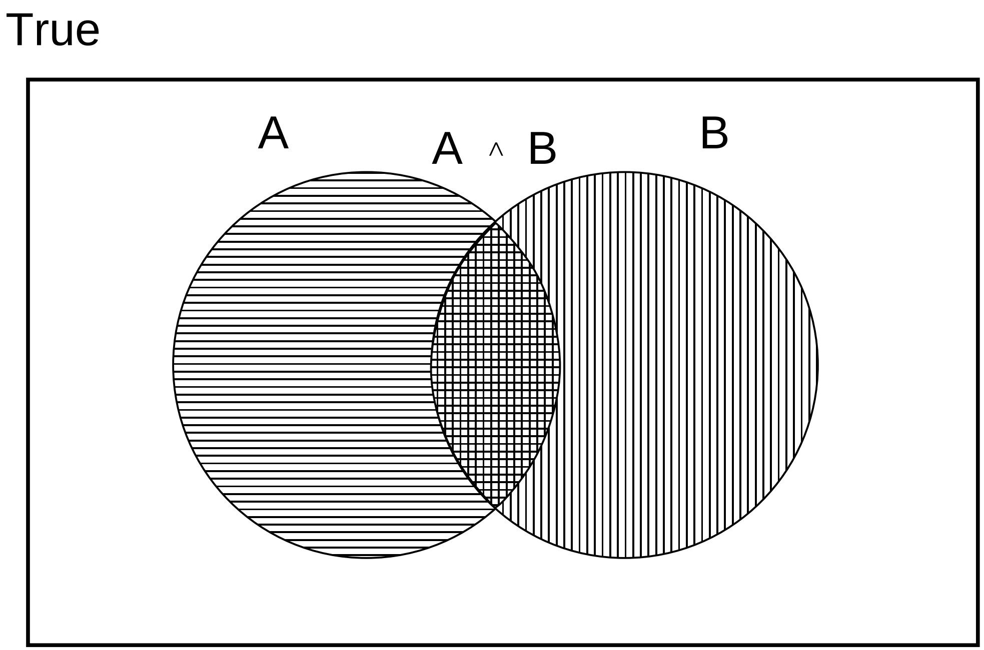
2. Valid propositions have probability 1

   $$P(\mathbf{True}) = P(\alpha \vee \neg\alpha) = 1$$

3. The probability of disjunction is defined as follows:

   $$P(\alpha \vee \beta) \quad = \quad P(\alpha) + P(\beta) - P(\alpha \wedge \beta)$$

# Understanding Axiom 3




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

# Conditional Probabilities

Conditional probabilities are defined in terms of unconditional ones

Whenever $P(B) > 0$,

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

This can be equivalently expressed as the product rule:

$$\begin{aligned} P(A \wedge B) &= P(A \mid B)\, P(B) \\ &= P(B \mid A)\, P(A) \end{aligned}$$

$A$ and $B$ are *independent* iff $P(A \mid B) = P(A)$
iff $P(B \mid A) = P(B)$
iff $P(A \wedge B) = P(A)P(B)$

# Random Variable

A *random variable* is a variable ranging over a certain domain of values

It is

- *discrete* if it ranges over a discrete (that is, countable) domain
- *continuous* if it ranges over the real numbers

*We will only consider discrete random variables with finite domains*

**Note:** Propositional variables can be seen as random variables over the Boolean domain

# Random Variables

| Variable | Domain |
|---|---|
| $Age$ | $\{1, 2, \ldots, 120\}$ |
| $Weather$ | $\{sunny, dry, cloudy, rain, snow\}$ |
| $Size$ | $\{small, medium, large\}$ |
| $Blonde$ | $\{true, false\}$ |

The probability that a random variable $X$ has value $val$ is written as

$$P(X = val)$$

**Note 1:** $P(A = true)$ is written shortly as $P(a)$ while $P(A = false)$ is written as $P(\neg a)$

**Note 2:** Traditionally, in Probability Theory variables are capitalized and constant values are not

# Probability Distribution

If $X$ is a random variable, we use the **bold case** $\mathbf{P}(X)$ to denote a *vector* of values for the probabilities of each individual element that $X$ can take.

Example:

$$\begin{aligned} P(Weather = sunny) &= 0.6 \\ P(Weather = rain) &= 0.2 \\ P(Weather = cloudy) &= 0.18 \\ P(Weather = snow) &= 0.02 \end{aligned}$$

Then $\mathbf{P}(Weather) = \langle 0.6, 0.2, 0.18, 0.02 \rangle$
(the value order of "sunny", "rain", cloudy", "snow" is assumed)

$\mathbf{P}(Weather)$ is called a *probability distribution* for the random variable $Weather$

# Joint Probability Distribution

If $X_1, \ldots, X_n$ are random variables,

$$\mathbf{P}(X_1, \ldots, X_n)$$

denotes their *joint probability distribution (JPD)*, an $n$-dimensional matrix specifying the probability of every possible combination of values for $X_1, \ldots, X_n$

Example

$$Sky: \quad \{sunny, cloudy, rain, snow\}$$
$$Wind: \quad \{true, false\}$$

$\mathbf{P}(Wind, Sky) =$

| | *sunny* | *cloudy* | *rain* | *snow* |
|---|---|---|---|---|
| *true* | 0.30 | 0.15 | 0.17 | 0.01 |
| *false* | 0.30 | 0.05 | 0.01 | 0.01 |

# Joint Probability Distribution

All relevant probabilities about a vector $\langle X_1, \ldots, X_n \rangle$ of random variables can be computed from $\mathbf{P}(X_1, \ldots, X_n)$

| | $S = sunny$ | $S = cloudy$ | $S = rain$ | $S = snow$ | $\mathbf{P}(W)$ |
|---|---|---|---|---|---|
| $W$ | 0.30 | 0.15 | 0.17 | 0.01 | 0.63 |
| $\neg W$ | 0.30 | 0.05 | 0.01 | 0.01 | 0.37 |
| $\mathbf{P}(S)$ | 0.60 | 0.20 | 0.18 | 0.02 | 1.00 |

$$
\begin{aligned}
P(S = rain \wedge W) &= 0.17 \\
P(S = rain) &= P(S = rain \wedge W) + P(S = rain \wedge \neg W) \\
&= 0.17 + 0.01 = 0.18 \\
P(W) &= 0.30 + 0.15 + 0.17 + 0.01 = 0.63 \\
P(S = rain \mid W) &= P(S = rain \wedge W)/P(W) \\
&= 0.17/0.63 = 0.27
\end{aligned}
$$

# Joint Probability Distribution

A joint probability distribution $\mathbf{P}(X_1, \ldots, X_n)$ provides complete information about the probabilities of its random variables.

However, JPD's are often hard to create (again because of incomplete knowledge of the domain).

Even when available, JPD tables are very expensive, or impossible, to store because of their size.

A JPD table for $n$ random variables, each ranging over $k$ distinct values, has $k^n$ entries!

A better approach is to come up with conditional probabilities as needed and compute the others from them.

# An Alternative to JPD: The Bayes Rule

Recall that for any fact $A$ and $B$,

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

From this we obtain the *Bayes Rule*:

$$P(B \mid A) \;=\; \frac{P(A \mid B)\,P(B)}{P(A)}$$

The rule is useful in practice because it is often easier to compute/estimate $P(A \mid B)$, $P(B)$, and $P(A)$ than to compute/estimate $P(B \mid A)$ directly

# Applying the Bayes Rule

What is the probability that a patient has meningitis ($M$) given that he has a stiff neck ($S$)?

$$P(m \mid s) \quad = \quad \frac{P(s \mid m)\, P(m)}{P(s)}$$

- $P(s \mid m)$ is easier to estimate than $P(m \mid s)$ because it refers to *causal knowledge*: meningitis typically causes stiff neck
- $P(s \mid m)$ can be estimated from past medical cases and the knowledge about how meningitis works
- Similarly, $P(m)$ and $P(s)$ can be estimated from statistical information

# Applying the Bayes Rule

The Bayes rule is helpful even in absence of (immediate) causal relationships

What is the probability that a blonde person ($B$) is Swedish ($S$)?

$$P(s \mid b) \;=\; \frac{P(b \mid s)P(s)}{P(b)}$$

All $P(b \mid s)$, $P(s)$, $P(b)$ are more easily estimated from statistical information

$$\begin{aligned}
P(b \mid s) &\approx \frac{\#\text{ of blonde Swedish}}{|\text{Swedish population}|} &= \frac{9}{10} \\
P(s) &\approx \frac{|\text{Swedish population}|}{|\text{world population}|} &= \ldots \\
P(b) &\approx \frac{\#\text{ of blondes}}{|\text{world population}|} &= \ldots
\end{aligned}$$

# Conditional Independence

In terms of exponential explosion, conditional probabilities do not seem any better than JPD's for computing the probability of a fact, given $n > 1$ pieces of evidence

$$P(meningitis \mid stiffNeck \wedge nausea \wedge \cdots \wedge doubleVision)$$

However, facts do not always depend on all the evidence.

Example:

$$P(meningitis \mid stiffNeck \wedge astigmatic) = P(meningitis \mid stiffNeck)$$

$Meningitis$ and $Astigmatic$ are *conditionally independent*, given knowledge of $StiffNeck$

# Conditional Probability Notation

Recall:

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

A general version holds for whole joint probability distributions, e.g.,

$$\mathbf{P}(Sky, Wind) = \mathbf{P}(Sky \mid Wind)\, \mathbf{P}(Wind)$$

stands for (with $S$ for $Sky$ and $W$ for $Wind$):

$$\begin{aligned}
P(S = sunny, W = true) &= P(S = sunny \mid W = true)\, P(W = true) \\
P(S = sunny, W = false) &= P(S = sunny \mid W = false)\, P(W = false) \\
&\vdots \\
P(S = snow, W = false) &= P(S = snow \mid W = false)\, P(W = false)
\end{aligned}$$

I.e., a $4 \times 2$ set of equations, not matrix multiplication

# Conditional Probability Notation

The *chain rule* is derived by successive application of product rule:

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

# Inference by enumeration

Let $\mathbf{X}$ be all the variables. Typically, we want

the posterior joint distribution of the *query variables* $\mathbf{Y}$

given specific values $\mathbf{e}$ for the *evidence variables* $\mathbf{E}$

Let the *hidden variables* be $\mathbf{H} = \mathbf{X} - \mathbf{Y} - \mathbf{E}$. Then,

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

where $\alpha = P(\mathbf{E}=\mathbf{e})^{-1}$

**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?

# Independence

$A$ and $B$ are *independent* iff

$$\mathbf{P}(A \mid B) = \mathbf{P}(A) \quad \text{iff} \quad \mathbf{P}(B \mid A) = \mathbf{P}(B) \quad \text{iff} \quad \mathbf{P}(A, B) = \mathbf{P}(A)\,\mathbf{P}(B)$$

Example:

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

32 entries reduced to 12; for $n$ independent biased coins, $2^n \rightarrow n$

**Problem:** Absolute independence is powerful but rare

For instance, dentistry is a large field with hundreds of variables, none of which are independent.

What to do? Exploit conditional independence

# Conditional independence

$\mathbf{P}(Toothache, Cavity, Catch)$ has $2^3 - 1 = 7$ independent entries

If I have a cavity, the probability that the probe catches in it doesn't depend on whether I have a toothache or not:

$$P(catch \mid toothache, cavity) = P(catch \mid cavity)$$

The same independence holds if I have no cavity:

$$P(catch \mid toothache, \neg cavity) = P(catch \mid \neg cavity)$$

$Catch$ is *conditionally independent* on $Toothache$ given $Cavity$:

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

Similarly:

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

# Conditional independence contd.

Write out full joint distribution using chain rule:

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

Now we have only $2 + 2 + 1 = 5$ independent entries

In most cases, conditional independence reduces the size of the representation of the JPD from exponential to linear in $n$

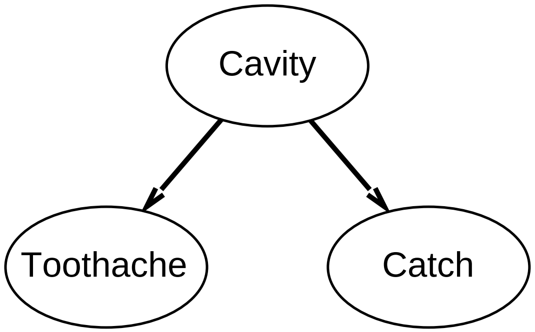
*Conditional independence is our most basic and robust form of knowledge about uncertain environments*

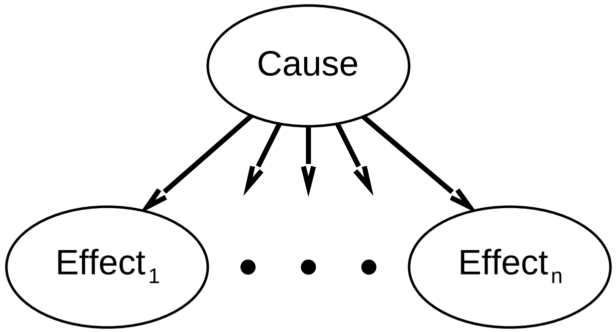
# Bayes' Rule and cond. independence

$$\mathbf{P}(Cavity \mid toothache \wedge catch)$$
$$= \alpha\, \mathbf{P}(toothache \wedge catch \mid Cavity)\, \mathbf{P}(Cavity)$$
$$= \alpha\, \mathbf{P}(toothache \mid Cavity)\, \mathbf{P}(catch \mid Cavity)\, \mathbf{P}(Cavity)$$

This is an example of a *naive Bayes* model:

$$\mathbf{P}(Cause, Effect_1, \ldots, Effect_n) = \Pi_{i=1}^{n} \mathbf{P}(Effect_i \mid Cause)\, \mathbf{P}(Cause)$$




Total number of parameters is linear in $n$

# Bayesian Networks

Exploiting conditional independence information is crucial in making (automated) probabilistic reasoning feasible

*Bayesian Networks* are a successful example of probabilistic systems that exploit conditional independence to reason efficiently under uncertainty