Machine Intelligence

4. Reasoning under Uncertainty, Part I: Basics

(Our Machinery for) Thinking About What is Likely to be True

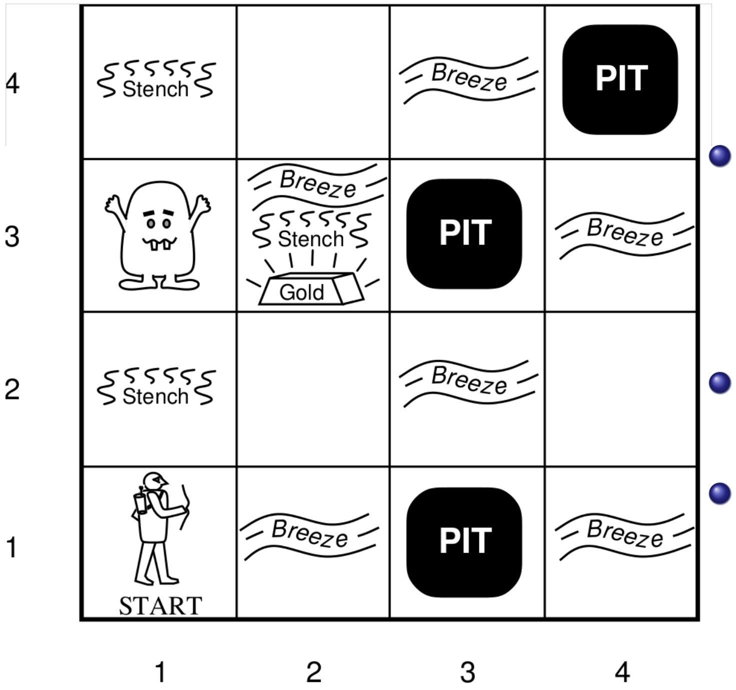
Álvaro Torralba

AALBORG UNIVERSITET

Fall 2022

Thanks to Thomas D. Nielsen and Jörg Hoffmann for slide sources

The Wumpus World



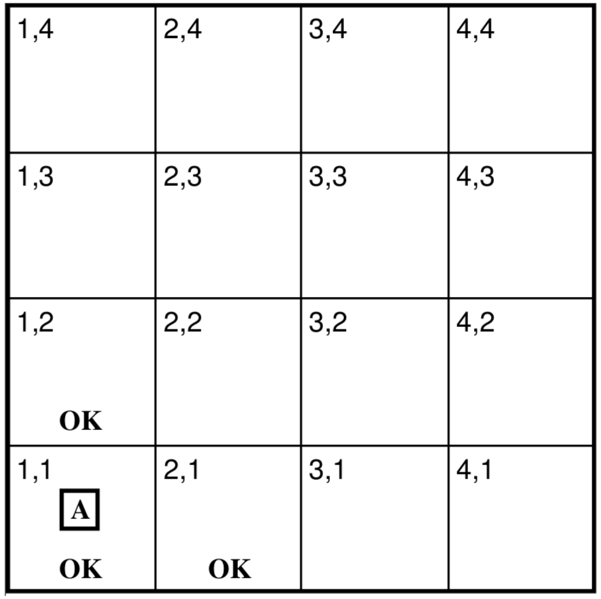
- **Actions:** *GoForward*, *TurnRight* (by 90°), *TurnLeft* (by 90°), *Grab* object in current cell, *Shoot* arrow in direction you're facing (you got exactly one arrow), *Leave* cave if you're in cell [1,1].

  → Fall down *Pit*, meet live *Wumpus*: Game Over.
- **Initial knowledge:** You're in cell [1,1] facing east. There's a Wumpus, and there's gold.
- **Goal**: Have the gold and be outside the cave.

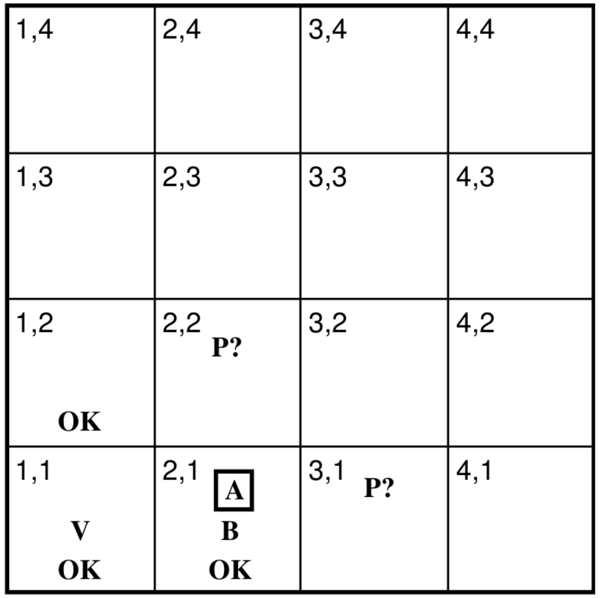
Percepts: $[Stench, Breeze, Glitter, Bump, Scream]$

Reasoning in the Wumpus World

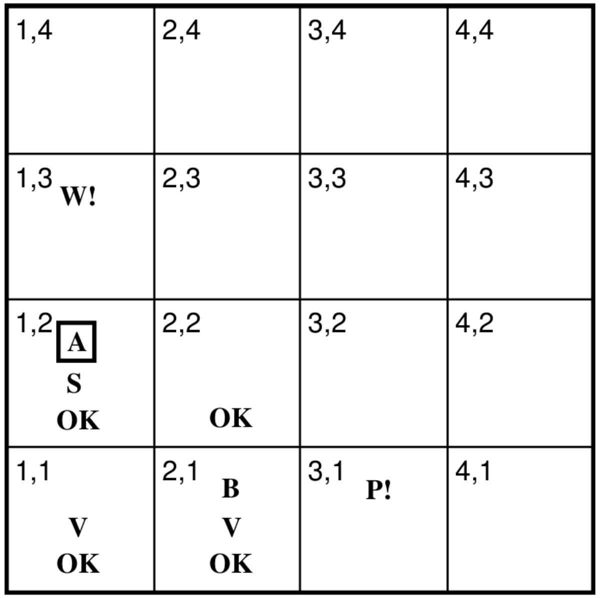
A: Agent, **V**: Visited, **OK**: Safe, **P**: Pit, **W**: Wumpus, **B**: Breeze, **S**: Stench, **G**: Gold



(1) Initial state



(2) One step to right



(3) Back, and up to [1,2]

→ The Wumpus is in [1,3]! How do we know?

→ There's a Pit in [3,1]! How do we know?

Agents that Think Rationally

Think Before You Act!

function KB-AGENT(*percept*) **returns** an *action*
  persistent: KB, a knowledge base
      t, a counter, initially 0, indicating time

  TELL(KB, MAKE-PERCEPT-SENTENCE(*percept*, t))
  action $\leftarrow$ ASK(KB, MAKE-ACTION-QUERY(t))
  TELL(KB, MAKE-ACTION-SENTENCE(*action*, t))
  $t \leftarrow t + 1$
  return *action*

→ "Thinking" = Reasoning about knowledge represented using logic.

Our Agenda for This Chapter

- **Propositional Logic:** What's the syntax and semantics? How can we capture deduction?
  → A brief introduction to logical reasoning.
- **Quantifying Uncertainty**: What to do when all we know is we don't know.
  → A bit of motivation for what comes next.
- **Unconditional Probabilities** and **Conditional Probabilities:** Which concepts and properties of probabilities will be used?
  → Mostly a recap of things you're familiar with from school.
- **Independence** and **Basic Probabilistic Reasoning Methods:** What simple methods are there to avoid enumeration and to deduce probabilities from other probabilities?
  → A basic tool set we'll need. (Still familiar from school?)
- **Bayes' Rule:** What's that "Bayes"? How is it used and why is it important?
  → The basic insight about how to invert the "direction" of conditional probabilities.
- **Conditional Independence:** How to capture and exploit complex relations between random variables?
  → Explains the difficulties arising when using Bayes' rule on multiple evidences. Conditional independence is used to ameliorate these difficulties.

Propositional Logic: Syntax

Atomic Propositions

Boolean variables are now seen as **atomic propositions**. Convention: start with lowercase letter.

Constraints	Logic
$A = \mathit{true}$	a
$A = \mathit{false}$	$\neg a$

Propositions (Formulas)

Using **logical connectives** more complex propositions are constructed:

$\neg p$	**not** p
$(p \wedge q)$	p **and** q
$(p \vee q)$	p **or** q
$(p \rightarrow q)$	p **implies** q

Example: "If **it rains** I'll **take my umbrella**, or I'll **stay home**"

$$\mathit{rains} \rightarrow (\mathit{umbrella} \vee \mathit{home})$$

Propositional Logic: Semantics I

An **interpretation** π for a set of atomic propositions $a_1, a_2, \ldots, a_n$ is an assignment of a truth value to each proposition (= possible world when atomic propositions seen as boolean variables):

$$\pi(a_i) \in \{\textit{true}, \textit{false}\}$$

An interpretation defines a truth value for all propositions:

$\pi(p)$	$\pi(\neg p)$
true	*false*
false	*true*

$\pi(p)$	$\pi(q)$	$\pi(p \wedge q)$
true	*true*	*true*
true	*false*	*false*
false	*true*	*false*
false	*false*	*false*

$\pi(p)$	$\pi(q)$	$\pi(p \vee q)$
true	*true*	*true*
true	*false*	*true*
false	*true*	*true*
false	*false*	*false*

Knowledge as Propositional Formulas

Satisfiability

A formula φ is:

- satisfiable if there exists I that satisfies φ.
- unsatisfiable if φ is not satisfiable.
- falsifiable if there exists I that doesn't satisfy φ.
- valid if $I \models \varphi$ holds for all I. We also call φ a tautology.

Knowledge Base

A Knowledge Base (KB) is a set of formulas that describe the agent's knowledge.

$\rightarrow$ Knowledge Base $=$ set of formulas, interpreted as a conjunction.

Definition (Model). A model of a a knowledge base KB is an interpretation I in which all the formulas in the knowledge base are true: $I \models \varphi$ for all $\varphi \in$ KB.

$\rightarrow$a model is a possible world that satisfies the constraint.

We denote by $M(\varphi)$ the set of all models of φ (i.e., the set of possible worlds where the formula is true).

Deduction

Deduction

deriving of a conclusion by reasoning

Remember (slide 4)? Does our knowledge of the cave entail a definite Wumpus position?

→ We don't know everything; what can we conclude from the things we *do* know?

Logical consequence (entailment)

Definition (Entailment). *Let Σ be a set of atomic propositions. We say that a set of formulas KB entails a formula φ, written $KB \models \varphi$, if φ is true in all models of KB, i.e., $M(\bigwedge_{\psi \in KB}) \subseteq M(\varphi)$. In this case, we also say that φ follows from KB.*

A formula φ is a **logical consequence** of a knowledge base *KB*, if every model of *KB* is a model of φ. Written:

$$KB \models g$$

(whenever *KB* is true, then φ also is true).

Example: $KB = \{man \to mortal, man\}$. Then

Simple Example

$$KB = \begin{cases} p \leftarrow q. \\ q. \\ r \leftarrow s. \end{cases}$$

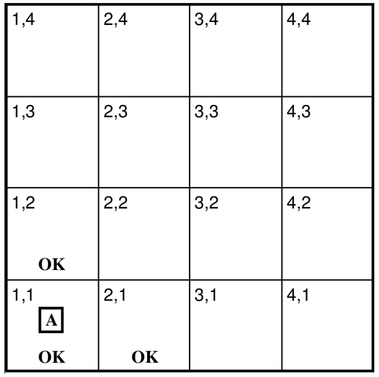
	p	q	r	s	Model?
I_1	*true*	*true*	*true*	*true*	is a model of KB
I_2	*false*	*false*	*false*	*false*	not a model of KB
I_3	*true*	*true*	*false*	*false*	is a model of KB
I_4	*true*	*true*	*true*	*false*	is a model of KB
I_5	*true*	*true*	*false*	*true*	not a model of KB

...

Which of p, q, r, q logically follow from KB?

$$KB \models p,\ KB \models q,\ KB \not\models r,\ KB \not\models s$$

Proof by Contradiction



Question!

Suppose that there exists an interpretation I in $M(KB)$ where the Wumpus is not at cell $(2,2)$. Can we conclude the cell $(2,2)$ is free?

(A): Yes (B): No

Uncertainty and Logic

Diagnosis: We want to build an expert dental diagnosis system, that deduces the cause (the disease) from the symptoms.

$\rightarrow$ Can we base this on logic?

Attempt 1: Say we have a toothache. How's about: $toothache \rightarrow cavity$

$\rightarrow$ Is this rule correct?

Attempt 2: So what about this: $toothache \rightarrow cavity \vee gum\_disease \vee \dots$

Attempt 3: Perhaps a *causal* rule is better? $cavity \rightarrow toothache$

- Is this rule correct?
- Does this rule allow to deduce a cause from a symptom?

Beliefs and Probabilities

What do we model with probabilities? Incomplete knowledge! We are not 100% sure, but we *believe to a certain degree* that something is true.

$\rightarrow$ Probability $\approx$ Our degree of belief, given our current knowledge.

Example (Diagnosis)

- $toothache \rightarrow cavity$ with 80% probability.
- But, for any given p, in reality we do, or do not, have cavity: 1 or 0!

$\rightarrow$ Probabilities represent and measure the uncertainty that stems from lack of knowledge.

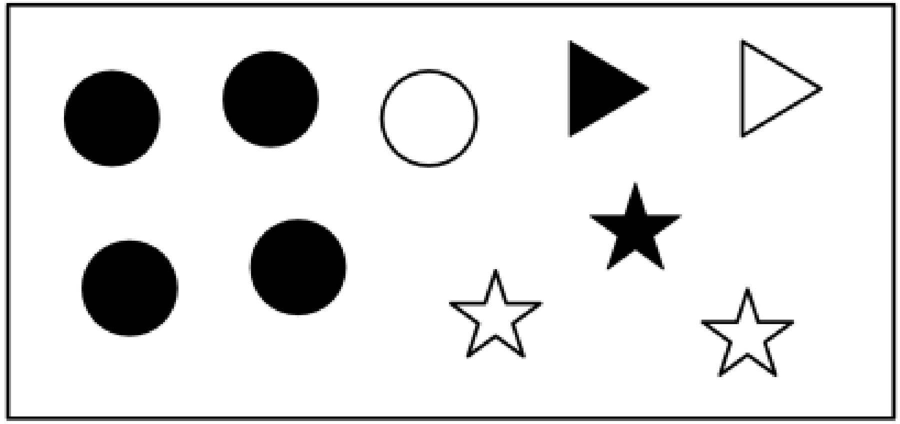
Probability Measures

Probability theory is built on the foundation of variables and worlds.

Worlds described by the variables:

- Filled: {true, false}
- Shape: {circle, triangle, star}

as well as position.



Probability measures

Ω: set of all possible worlds (for a given, fixed set of variables). A **probability measure over** Ω, is a function P, that assigns **probability values**

$$P(\Omega') \in [0, 1]$$

to subsets $\Omega' \subseteq \Omega$, such that

Axiom 1: $P(\Omega) = 1$.

Axiom 2: if $\Omega_1 \cap \Omega_2 = \emptyset$, then $P(\Omega_1 \cup \Omega_2) = P(\Omega_1) + P(\Omega_2)$.

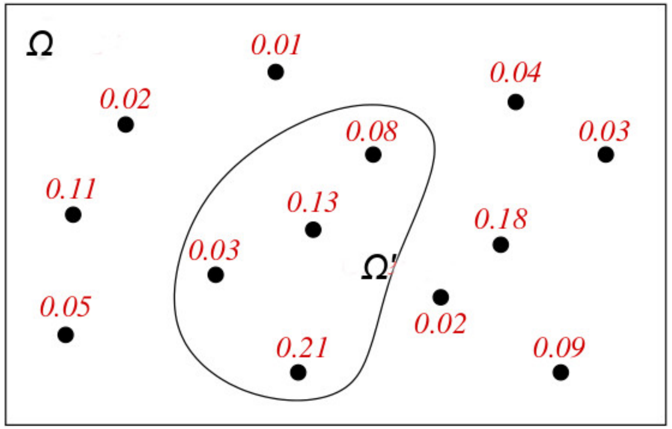
Simplification for finite Ω

If all variables have a finite domain, then

- Ω is finite, and
- a probability distribution is defined by assigning a probability value $P(\omega)$ to each individual possible world $\omega \in \Omega$.

For any $\Omega' \subseteq \Omega$ then

$$P(\Omega') = \sum_{\omega \in \Omega'} P(\omega)$$



$$P(\Omega') = 0.08 + 0.13 + 0.03 + 0.21 = 0.45$$

Note: In general, random variables can have arbitrary domains. Here, we consider finite-domain random variables only, and Boolean random variables most of the time.

Random Variables and Distributions

Definition (Random Variables). Variables defining possible worlds on which probabilities are defined are called random variables.

Distributions

For a random variable A, and $a \in D_A$ we have the probability

$$P(A = a) = P(\{\omega \in \Omega \mid A = a \text{ in } \omega\})$$

The **probability distribution of** A (written $P(A)$) is the function on D_A that maps each a to its probability $P(A = a)$.

Example:

$$\mathbf{P}(Headache) = \langle F \mapsto 0.1, T \mapsto 0.9 \rangle$$

$$\mathbf{P}(Weather) = \langle sunny \mapsto 0.7, rain \mapsto 0.2, cloudy \mapsto 0.08, snow \mapsto 0.02 \rangle$$

Joint Probability Distributions

Extension to several random variables:

$$P(A_1, \dots, A_k)$$

is the joint distribution of $A_1, \dots, A_k$. The joint distribution maps tuples $(a_1, \dots, a_k)$ with $a_i \in D_{A_i}$ to the probability

$$P(A_1 = a_1, \dots, A_k = a_k)$$

Example: $\mathbf{P}(Headache, Weather) =$

	$Headache = true$	$Headache = false$
$Weather = sunny$	$P(W = sunny \wedge headache)$	$P(W = sunny \wedge \neg headache)$
$Weather = rain$		
$Weather = cloudy$		
$Weather = snow$		

Terminology:

- Given random variables $\{X_1, \dots, X_n\}$, an atomic event (world) is an assignment of values to all variables.
- Given random variables $\{X_1, \dots, X_n\}$, the full joint probability distribution, denoted $\mathbf{P}(X_1, \dots, X_n)$, lists the probabilities of all atomic events.

$\rightarrow$ All worlds are disjoint (their pairwise conjunctions all are $\bot$); the sum of all fields is 1 (corresponds to their disjunction $\top$).

Probabilities of propositions

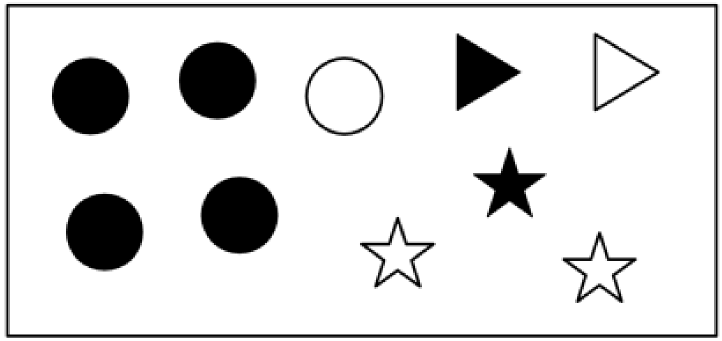
A probability distribution over possible worlds defines probabilities for formulas φ:

$$P(\alpha) = \sum_{\omega \in \Omega : \omega \in M(\varphi)} P(\omega)$$

→ Propositions represent sets of atomic events: the interpretations satisfying the formula.

Notation: Instead of $P(a \wedge b)$, we often write $P(a, b)$.

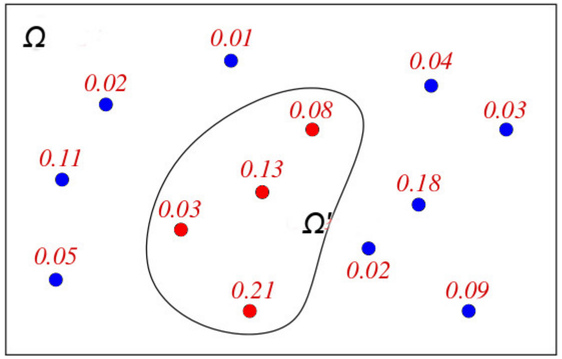
Example



Assume probability for each world is 0.1:

- $P(\mathit{Shape} = \mathit{circle}) =$
- $P(\mathit{Filled} = \mathit{false}) =$
- $P(\mathit{Shape} = c \wedge \mathit{Filled} = f) =$

Another example



Basic probability axioms

Axiom

If $\mathcal{A}$ and $\mathcal{B}$ are disjoint, then $P(\mathcal{A} \cup \mathcal{B}) = P(\mathcal{A}) + P(\mathcal{B})$.

Example

Consider a deck with 52 cards. If $\mathcal{A} = \{2, 3, 4, 5\}$ and $\mathcal{B} = \{7, 8\}$, then

$$P(\mathcal{A} \cup \mathcal{B}) = P(\mathcal{A}) + P(\mathcal{B}) = \frac{4}{13} + \frac{2}{13} = \frac{6}{13}.$$

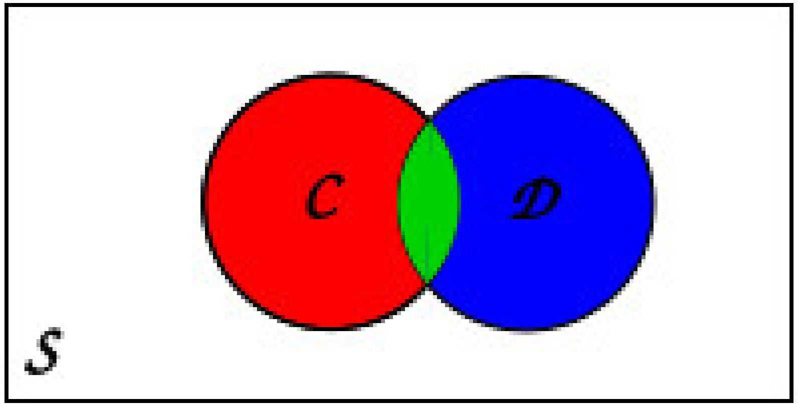
More generally

If $\mathcal{C}$ and $\mathcal{D}$ are not disjoint, then $P(\mathcal{C} \cup \mathcal{D}) = P(\mathcal{C}) + P(\mathcal{D}) - P(\mathcal{C} \cap \mathcal{D})$.

Example

If $\mathcal{C} = \{2, 3, 4, 5\}$ and $\mathcal{D} = \{\spadesuit\}$, then

$$P(\mathcal{C} \cup \mathcal{D}) = \frac{4}{13} + \frac{1}{4} - \frac{4}{52} = \frac{25}{52}.$$



Updating Your Beliefs

→ Do probabilities change as we gather new knowledge?

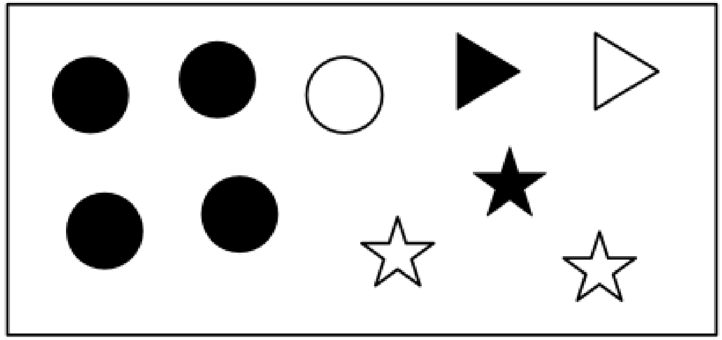
Conditional Probabilities

Definition. *Given propositions p and e where $P(e) \neq 0$, the conditional probability, or posterior probability, of p given e, written $P(p \mid e)$, is defined as:*

$$P(p \mid e) = \frac{P(p \wedge e)}{P(e)}$$

$\rightarrow$ The likelihood of having p and e, within the set of outcomes where we have e.

Example

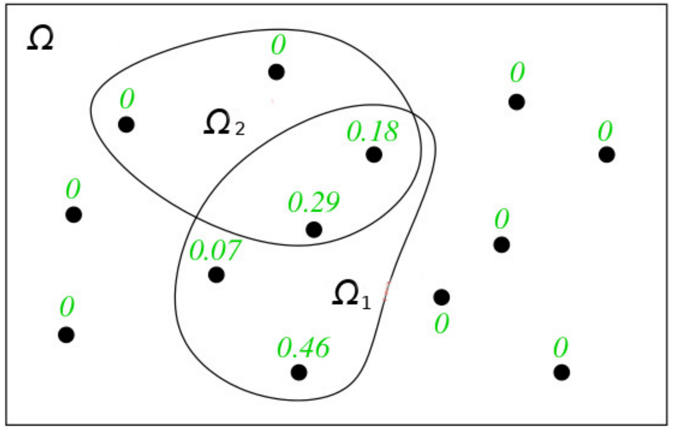


(probability for each world is 0.1)

$$P(S{=}circ. \mid Fill = f) = \frac{P(S{=}circ. \wedge Fill = f)}{P(Fill = f)} = \frac{0.1}{0.4} = 0.25E$$

What is the probability of $P(S{=}star \mid Fill = f)$?

Another example



- e and p are represented by possible worlds Ω_1 and Ω_2
- division by $P(\Omega_1)$ already in green numbers

$$P(\Omega_2 \mid \Omega_1) = 0.18 + 0.29$$

Conditional Probability Distributions

Definition. *Given random variables X and Y, the conditional probability distribution of X given Y, written $\mathbf{P}(X \mid Y)$, is the table of all conditional probabilities of values of X given values of Y.*

$\rightarrow$ For sets of variables: $\mathbf{P}(X_1, \dots, X_n \mid Y_1, \dots, Y_m)$.

Example: $\mathbf{P}(Weather \mid Headache) =$

	$Headache = true$	$Headache = false$
$Weather = sunny$	$P(W = sunny \mid headache)$	$P(W = sunny \mid \neg headache)$
$Weather = rain$		
$Weather = cloudy$		
$Weather = snow$		

$\rightarrow$ "The probability of sunshine given that I have a headache?"

Working with the Full Joint Probability Distribution

Example:

	$toothache$	$\neg toothache$
$cavity$	0.12	0.08
$\neg cavity$	0.08	0.72

$\rightarrow$ How to compute $P(cavity)$?

$\rightarrow$ How to compute $P(cavity \vee toothache)$?

$\rightarrow$ How to compute $P(cavity \mid toothache)$?

$\rightarrow$ All relevant probabilities can be computed using the full joint probability distribution, by expressing propositions as disjunctions of atomic events.

Working with the Full Joint Probability Distribution??

→ So, is it a good idea to use the full joint probability distribution?

→ So, is there a compact way to represent the full joint probability distribution? Is there an efficient method to work with that representation?

→ Not in general, but it works in many cases. We can work directly with conditional probabilities, and exploit (conditional) independence.

→ Bayesian networks. (First, we do the simple case.)

Independence

Definition. *Events a and b are independent if $P(a \wedge b) = P(a)P(b)$.*

Proposition. *Given independent events a and b where $P(b) \neq 0$, we have $P(a \mid b) = P(a)$.*

Proof. By definition, $P(a \mid b) = \frac{P(a \wedge b)}{P(b)}$,

Examples:

- $P(\mathit{Dice1} = 6 \wedge \mathit{Dice2} = 6) = 1/36$.
- $P(W = \mathit{sunny} \mid \mathit{headache}) = P(W = \mathit{sunny})$ unless you're weather-sensitive (cf. slide 26).
- But *toothache* and *cavity* are NOT independent. The fraction of "*cavity*" is higher within "*toothache*" than within "$\neg$*toothache*". $P(\mathit{toothache}) = 0.2$ and $P(\mathit{cavity}) = 0.2$, but $P(\mathit{toothache} \wedge \mathit{cavity}) = 0.12 > 0.04$.

Definition. *Random variables X and Y are independent if $\mathbf{P}(X, Y) = \mathbf{P}(X)\mathbf{P}(Y)$.* (System of equations!)

Example: Football statistics

Results for Bayern Munich and SC Freiburg in seasons 2001/02 and 2003/04. (Not counting the matches Munich vs. Freiburg):

$$D_{Munich)} = D_{Freiburg} = \{Win,\ Draw,\ Loss\}$$

	2001/02
Munich:	`LWDWWWWWWWLDLDLDLWLDWWWDWDDWWWW`
Freiburg:	`WLLDDWLDWDWLLLDDLWDDLLDLLLLLLWLW`
	2003/04
Munich:	`WDWWLDWWDWLWWDDWDWLWWWDDWWWLWWLL`
Freiburg:	`LDDWDWLWLLLWWLWLWLLDWLDDWDLLLWLD`

Summary:

	Freiburg			
Munich	W	D	L	
W	12	9	15	36
D	3	4	9	16
L	6	4	2	12
	21	17	26	

Independence of Outcomes

The joint distribution of *Munich* and *Freiburg*:

$P(\mathit{Munich},\mathit{Freiburg})$:

Munich	*Freiburg* W	D	L	$P(\mathit{Munich})$
W	.1875	.1406	.2344	.5625
	.571	.529	.577	
D	.0468	.0625	.1406	.25
	.143	.235	.346	
L	.0937	.0625	.0312	.1875
	.285	.235	.077	
$P(\mathit{Freiburg})$	.3281	.2656	.4062	

Conditional distribution: $P(\mathit{Munich} \mid \mathit{Freiburg})$

We have (almost):

$$P(\mathit{Munich} \mid \mathit{Freiburg}) = P(\mathit{Munich})$$

The variables *Munich* and *Freiburg* are **independent**.

Independent Variables

Definition of Independence

The variables $A_1, \dots, A_k$ and $B_1, \dots, B_m$ are **independent** if

$$P(A_1, \dots, A_k \mid B_1, \dots, B_m) = P(A_1, \dots, A_k)$$

This is equivalent to:

$$P(B_1, \dots, B_m \mid A_1, \dots, A_k) = P(B_1, \dots, B_m)$$

and also to:

$$P(A_1, \dots, A_k, B_1, \dots, B_m) = P(A_1, \dots, A_k) \cdot P(B_1, \dots, B_m)$$

Compact Specifications by Independence

Independence properties can greatly simplify the specification of a joint distribution:

$M =$	$F =$ W	D	L	$P(M)$
W				.5625
D	M and F are independent			.25
L				.1875
$P(F)$	.3281	.2656	.4062	

The probability for each possible world then is defined, e.g.

$$P(M = D, F = L) = 0.25 \cdot 0.4062 = 0.10155$$

$\rightarrow$ Independence can be exploited to represent the full joint probability distribution more compactly.

$\rightarrow$ Usually, random variables are independent only under particular conditions: conditional independence, see later.

The Product Rule

Proposition (Product Rule). *Given propositions A and B, $P(a \wedge b) = P(a \mid b)P(b)$.* (Direct from definition.)

Example: $P(cavity \wedge toothache) = P(toothache \mid cavity)P(cavity)$.

→ If we know the values of $P(a \mid b)$ and $P(b)$, then we can compute $P(a \wedge b)$.

→ Similarly, $P(a \wedge b) = P(b \mid a)P(a)$.

Notation: $\mathbf{P}(X, Y) = \mathbf{P}(X \mid Y)\mathbf{P}(Y)$ is a system of equations:

$$\begin{aligned} P(W = sunny \wedge headache) &= P(W = sunny \mid headache)P(headache) \\ P(W = rain \wedge headache) &= P(W = rain \mid headache)P(headache) \\ \ldots &= \ldots \\ P(W = snow \wedge \neg headache) &= P(W = snow \mid \neg headache)P(\neg headache) \end{aligned}$$

→ Similar for unconditional distributions, $\mathbf{P}(X, Y) = \mathbf{P}(X)\mathbf{P}(Y)$.

The Chain Rule

Proposition (Chain Rule). *Given random variables $X_1, \dots, X_n$, we have* $\mathbf{P}(X_1, \dots, X_n) = \mathbf{P}(X_n \mid X_{n-1}, \dots, X_1) * \mathbf{P}(X_{n-1} \mid X_{n-2}, \dots, X_1) * \dots * \mathbf{P}(X_2 \mid X_1) * \mathbf{P}(X_1)$.

Example: $P(\neg brush \wedge cavity \wedge toothache)$
$= P(toothache \mid cavity, \neg brush)P(cavity, \neg brush)$
$= P(toothache \mid cavity, \neg brush)P(cavity \mid \neg brush)P(\neg brush)$.

Proof. Iterated application of Product Rule.

Note: This works *for any ordering* of the variables.

→ We can recover the probability of atomic events from sequenced conditional probabilities for any ordering of the variables.

→ First of the four basic techniques in Bayesian networks.

Marginalization

→ Extracting a sub-distribution from a larger joint distribution:

Proposition (Marginalization). *Given sets* $\mathbf{X}$ *and* $\mathbf{Y}$ *of random variables, we have:*

$$\mathbf{P}(\mathbf{X}) = \sum_{\mathbf{y} \in \mathbf{Y}} \mathbf{P}(\mathbf{X}, \mathbf{y})$$

where $\sum_{\mathbf{y} \in \mathbf{Y}}$ *sums over all possible value combinations of* $\mathbf{Y}$.

Example: (Note: Equation system!)

$$\mathbf{P}(Cavity) = \sum_{y \in Toothache} \mathbf{P}(Cavity, y)$$

$$P(cavity) = P(cavity, toothache) + P(cavity, \neg toothache)$$

$$P(\neg cavity) = P(\neg cavity, toothache) + P(\neg cavity, \neg toothache)$$

Questionnaire

Question!

Say $P(dog) = 0.4$, $\neg dog \leftrightarrow cat$, **and** $P(likeslasagna \mid cat) = 0.5$. **Then** $P(likeslasagna \wedge cat) =$

(A): 0.2 (B): 0.5

(C): 0.475 (D): 0.3

Question!

Can we compute the value of $P(likeslasagna)$, **given the above informations?**

(A): Yes. (B): No.

Normalization: Idea

Problem: We know $P(cavity \wedge toothache)$ but don't know $P(toothache)$:

$$P(cavity \mid toothache) = \frac{P(cavity \wedge toothache)}{P(toothache)} = \frac{0.12}{P(toothache)}$$

Step 1: Case distinction over the values of $Cavity$:

$$P(\neg cavity \mid toothache) = \frac{P(\neg cavity \wedge toothache)}{P(toothache)} = \frac{0.08}{P(toothache)}$$

Step 2: Assuming placeholder $\alpha := 1/P(toothache)$:

$$P(cavity \mid toothache) = \alpha P(cavity \wedge toothache) = \alpha 0.12$$
$$P(\neg cavity \mid toothache) = \alpha P(\neg cavity \wedge toothache) = \alpha 0.08$$

Step 3: Fixing $toothache$ to be true, view $P(cavity \wedge toothache)$ vs. $P(\neg cavity \wedge toothache)$ as the relative weights of $P(cavity)$ vs. $P(\neg cavity)$ within $toothache$. Then normalize their summed-up weight to 1:

$$1 = \alpha(0.12 + 0.08) \Rightarrow \alpha = 1/(0.12 + 0.08) = 1/0.2 = 5$$

$\rightarrow$ α is the normalization constant scaling the sum of relative weights to 1.

Normalization: Formal

Definition. *Given a vector* $\langle w_1, \dots, w_k \rangle$ *of numbers in* $[0,1]$ *where* $\sum_{i=1}^{k} w_i \leq 1$, *the normalization constant* α *is* $\alpha\langle w_1, \dots, w_k \rangle := 1/\sum_{i=1}^{k} w_i$.

Example: $\alpha\langle 0.12, 0.08 \rangle = 5\langle 0.12, 0.08 \rangle = \langle 0.6, 0.4 \rangle$.

Proposition (Normalization). *Given a random variable* X *and an event* $\mathbf{e}$, *we have* $\mathbf{P}(X \mid \mathbf{e}) = \alpha\mathbf{P}(X, \mathbf{e})$.

Proof.

Example: $\alpha\langle P(cavity \wedge toothache), P(\neg cavity \wedge toothache)\rangle = \alpha\langle 0.12, 0.08 \rangle$, so $P(cavity \mid toothache) = 0.6$, and $P(\neg cavity \mid toothache) = 0.4$.

Normalization+Marginalization: Given "query variable" X, "observed event" $\mathbf{e}$, and "hidden variables" set $\mathbf{Y}$: $\mathbf{P}(X \mid \mathbf{e}) = \alpha\mathbf{P}(X, \mathbf{e}) = \alpha \sum_{\mathbf{y} \in \mathbf{Y}} \mathbf{P}(X, \mathbf{e}, \mathbf{y})$.

$\rightarrow$ Second of the four basic techniques in Bayesian networks.

Questionnaire

Question!

Say we know $P(likeschappi \wedge dog) = 0.32$ **and** $P(\neg likeschappi \wedge dog) = 0.08$**. Can we compute** $P(likeschappi \mid dog)$**?**

(A): Yes. (B): No.

Bayes' Rule

Proposition (Bayes' Rule). *Given propositions A and B where $P(a) \neq 0$ and $P(b) \neq 0$, we have:*

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

Proof. By definition, $P(a \mid b) = \frac{P(a \wedge b)}{P(b)}$

Notation: (System of equations)

$$\mathbf{P}(X \mid Y) = \frac{\mathbf{P}(Y \mid X)\mathbf{P}(X)}{\mathbf{P}(Y)}$$

Applying Bayes' Rule

Example: Say we know that $P(toothache \mid cavity) = 0.6$, $P(cavity) = 0.2$, and $P(toothache) = 0.2$.

$\rightarrow$ We can compute $P(cavity \mid toothache)$:

Ok, but: Why don't we simply assess $P(cavity \mid toothache)$ directly?

- $P(toothache \mid cavity)$ is causal, $P(cavity \mid toothache)$ is diagnostic.
- Causal dependencies are robust over frequency of the causes.
  $\rightarrow$ Example: If there is a cavity epidemic then $P(cavity \mid toothache)$ increases, but $P(toothache \mid cavity)$ remains the same.
- Also, causal dependencies are often easier to assess.

$\rightarrow$ Bayes' rule allows to perform diagnosis (observing a symptom, what is the cause?) based on prior probabilities and causal dependencies.

Questionnaire

Question!

Say $P(dog) = 0.4$**,** $P(likeschappi \mid dog) = 0.8$**, and** $P(likeschappi) = 0.5$**. What is** $P(dog \mid likeschappi)$**?**

(A): 0.8 (B): 0.64

(C): 0.9 (D): 0.32

Bayes' rule for variables

Bayes' rule can also be written for variables:

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

Example

Consider the variables

- Temp : $\mathrm{sp}(\text{Temp}) = \{l, m, h\}$
- Sensor : $\mathrm{sp}(\text{Sensor}) = \{l, m, h\}$

$P(\text{Temp}) = (red0.1, green0.6, lightblue0.3)$

$P(\text{Sensor}|\text{Temp}) =$

		Temp		
		l	m	h
Sensor	l	red0.8	green0.1	lightblue0.05
	~~m~~	~~0.15~~	~~0.8~~	~~0.1~~
	~~h~~	~~0.05~~	~~0.1~~	~~0.85~~

Assume we observe $S = \text{low}$:

$P(T|\text{low}) = \frac{P(\text{low}|T)P(T)}{\sum_T P(\text{low}|T)P(T)} =$

	Temp		
	l	m	
$S = \text{low}$	$red0.08/yellow0.155$	$green0.06/yellow0.155$	$lightbl$

	Temp		
	l	m	h
$S = \text{low}$	0.52	0.39	0.09

$P(\text{low}|T)P(T) =$

	Temp		
	l	m	h
$S = \text{low}$	red0.08	green0.06	lightblue0.015

$\rightarrow$ $P(S =$ yellow0.

Questionnaire

Conditional Independence

Definition. *Given sets of random variables* $\mathbf{Z}_1$, $\mathbf{Z}_2$, $\mathbf{Z}$, *we say that* $\mathbf{Z}_1$ *and* $\mathbf{Z}_2$ *are conditionally independent given* $\mathbf{Z}$ *if:*

$$\mathbf{P}(\mathbf{Z}_1, \mathbf{Z}_2 \mid \mathbf{Z}) = \mathbf{P}(\mathbf{Z}_1 \mid \mathbf{Z})\mathbf{P}(\mathbf{Z}_2 \mid \mathbf{Z})$$

We alternatively say that $\mathbf{Z}_1$ *is conditionally independent of* $\mathbf{Z}_2$ *given* $\mathbf{Z}$.

Note: The definition is symmetric regarding the roles of $\mathbf{Z}_1$ and $\mathbf{Z}_2$: $hairlength$ is conditionally independent of $height$, *and vice versa*.

Bayes' Rule with Multiple Evidence

Example: Say we know from medicinical studies that $P(cavity) = 0.2$, $P(toothache \mid cavity) = 0.6$, $P(toothache \mid \neg cavity) = 0.1$, $P(catch \mid cavity) = 0.9$, and $P(catch \mid \neg cavity) = 0.2$. Now, in case we did observe the symptoms toothache and catch (the dentist's probe catches in the aching tooth), what would be the likelihood of having a cavity? What is $P(cavity \mid catch, toothache)$?

By Bayes' rule we get:

$$P(cavity \mid catch, toothache) = \frac{P(catch, toothache \mid cavity)P(cavity)}{P(catch, toothache)}$$

Question!

So, is everything fine? Do we just need some more medicinical studies?

(A): Yes. (B): No.

Bayes' Rule with Multiple Evidence, ctd.

Second attempt: First Normalization (slide 42), then Chain Rule (slide 38) using ordering $X_1 = Cavity, X_2 = Catch, X_3 = Toothache$:

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

Close, but no Banana: Less red (i.e.unknown) probabilities, but still $\mathbf{P}(toothache \mid catch, Cavity)$.

But: Are $Toothache$ and $Catch$ independent?

Conditional Independence, ctd.

Proposition. *If $\mathbf{Z}_1$ and $\mathbf{Z}_2$ are conditionally independent given $\mathbf{Z}$, then* $\mathbf{P}(\mathbf{Z}_1 \mid \mathbf{Z}_2, \mathbf{Z}) = \mathbf{P}(\mathbf{Z}_1 \mid \mathbf{Z})$.

Proof. By definition, $\mathbf{P}(\mathbf{Z}_1 \mid \mathbf{Z}_2, \mathbf{Z}) = \frac{\mathbf{P}(\mathbf{Z}_1, \mathbf{Z}_2, \mathbf{Z})}{\mathbf{P}(\mathbf{Z}_2, \mathbf{Z})}$

Example: Using $\{Toothache\}$ as $\mathbf{Z}_1$, $\{Catch\}$ as $\mathbf{Z}_2$, and $\{Cavity\}$ as $\mathbf{Z}$:
$\mathbf{P}(Toothache \mid Catch, Cavity) = \mathbf{P}(Toothache \mid Cavity)$.

$\rightarrow$ In the presence of conditional independence, we can drop variables from the right-hand side of conditional probabilities.

$\rightarrow$ Third of the four basic techniques in Bayesian networks. Last missing technique: "Capture variable dependencies in a graph"; illustration see Conclusions, details see **Next Chapter**.

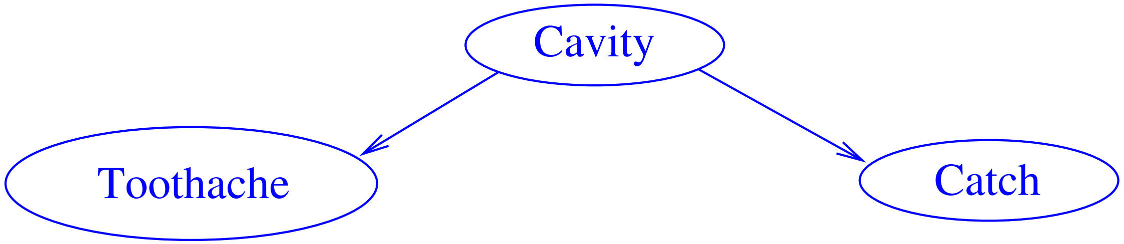
Summary

- Reasoning can be attained by a combination of logic and probability.
- Deduction is about deriving conclusions that follow logically from our knowledge base.
- Uncertainty is unavoidable in many environments, namely whenever agents do not have perfect knowledge.
- Probabilities express the degree of belief of an agent, given its knowledge, into an event.
- Conditional probabilities express the likelihood of an event given observed evidence.
- Assessing a probability means to use statistics to approximate the likelihood of an event.
- Bayes' rule allows us to derive, from probabilities that are easy to assess, probabilities that aren't easy to assess.
- Given multiple evidence, we can exploit conditional independence.

  $\rightarrow$ Bayesian networks (up next) do this, in a comprehensive manner (see next slides for some spoilers of where are we headed).

Exploiting Conditional Independence: Overview

1. **Graph captures variable dependencies:** (Variables $X_1, \dots, X_n$)



$\rightarrow$ Given evidence $\mathbf{e}$, want to know $\mathbf{P}(X \mid e)$. Remaining vars: $\mathbf{Y}$.

2. **Normalization+Marginalization:**

$$\mathbf{P}(X \mid \mathbf{e}) = \alpha \mathbf{P}(X, \mathbf{e}); \text{ if } \mathbf{Y} \neq \emptyset \text{ then } \mathbf{P}(X \mid \mathbf{e}) = \alpha \sum_{\mathbf{y} \in \mathbf{Y}} \mathbf{P}(X, \mathbf{e}, \mathbf{y})$$

$\rightarrow$ A sum over atomic events!

3. **Chain rule:** Order $X_1, \dots, X_n$ consistently with dependency graph.

$$\mathbf{P}(X_1, \dots, X_n) = \mathbf{P}(X_n \mid X_{n-1}, \dots, X_1) * \mathbf{P}(X_{n-1} \mid X_{n-2}, \dots, X_1) * \dots * \mathbf{P}(X_1)$$

4. **Exploit conditional independence:** Instead of $\mathbf{P}(X_i \mid X_{i-1}, \dots, X_1)$, with previous slide we can use $\mathbf{P}(X_i \mid Parents(X_i))$.

$\rightarrow$ Bayesian networks!

Exploiting Conditional Independence: Example