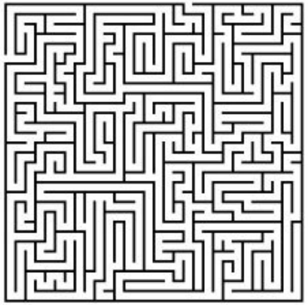
Uncertainty and Probability

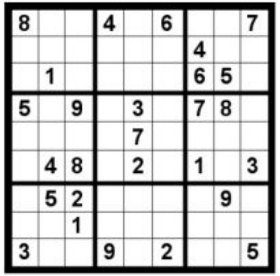
Lecture outline

- Uncertainty
- Probability
  - Combinatorics
  - Joint probability
  - Marginal probability
  - Conditional probability
  - Independence
  - Conditional independence
  - Bayes' rule

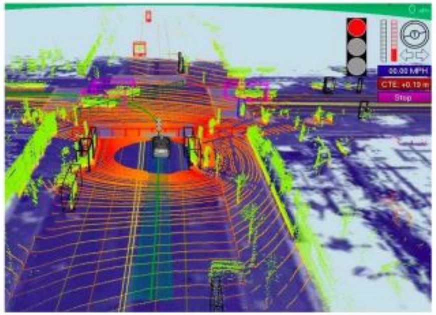
Uncertainty

Deterministic Reasoning vs Probabilistic Reasoning






vs.




Uncertainty

- If I take my umbrella, I won't get wet in the rain
  - **Deterministic reasoning**
- But there is a chance that:
  - The umbrella might brake and I'll get wet
  - There will be strong winds spraying water everywhere and I'll get wet
  - **Probabilistic reasoning**
- Agents almost never have perfect information about the world
  - In those situations, the agent must reason under **uncertainty**
  - Uncertainty can also arise because of the agent's incorrect/incomplete understanding of its environment

Making decisions under uncertainty

Let **Action** $\mathbf{A_t}$ = "*leave for airport t minutes before the flight*"

Will $\mathbf{A_t}$ get me to the flight in time?

Problems:

- Partial observability (road state, other drivers,...)
- Noisy sensors (unreliable traffic reports, weather reports,...)
- Uncertainty in action outcomes (flat tyre, mechanical failure,...)
- Complexity of modeling and predicting traffic

Making decisions under uncertainty

A purely logical approach either

1. Risks falsehood
    a. A_{25} will get me there on time, or
2. Leads to conclusions that are too weak for decision making
    a. A_{25} will get me there on time if there is no accident on the way and there is no rain and there is no problem with the car, and …

A_{2500} might get me there in time for the flight but I might have to stay at the airport for far too long!

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 First Order Logic 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 time
- 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

Methods for handling uncertainty

- Default or non-monotonic logic:
  - E.g., assume $\mathbf{A_{25}}$ works unless contradicted by evidence
  - Issues: What assumptions are reasonable? How to handle contradiction?
- Probability
  - E.g., Given the available evidence, A_{25} will get me there on time with **probability 0.04**

Making decisions under uncertainty

- Suppose the agent believes the following:
  - P(A25 gets me there on time) = 0.04
  - P(A90 gets me there on time) = 0.70
  - P(A120 gets me there on time) = 0.95
  - P(A1440 gets me there on time) = 0.9999
- Which action should the agent choose?
  - Depends on preferences for missing flight vs. time spent waiting
  - Encapsulated by a **utility function**
    - Attempts to quantify satisfaction/happiness
- The agent should choose the action that maximizes the **expected utility**:
  - $P(A_t \text{ succeeds}) * U(A_t \text{ succeeds}) + P(A_t \text{ fails}) * U(A_t \text{ fails})$

Making decisions under uncertainty

- More generally: the ***expected utility of an action*** is defined as:
- $EU(a) = \Sigma_{\text{outcomes of a}} P(\text{outcome} \mid a) * U(\text{outcome})$
- Utility theory is used to represent and infer preferences
- Decision theory = probability theory + utility theory

Probability

Probability

- Probabilistic assertions summarize effects of
  - laziness: failure to enumerate exceptions, etc.
  - ignorance: lack of relevant facts, initial conditions, etc.
- Subjective or Bayesian probability:
  - Probabilities relate propositions to one's own state of knowledge
  - e.g., $P(A_{25}$ gets me there on time | no reported accidents$) = 0.06$
- Probabilities of propositions change with new evidence:
  - E.g., $P(A_{25}$ gets me there on time | no reported accidents, 5 a.m.$) = 0.15$

Probability basics

- Begin with a set **Ω** – the sample space
  - e.g., 6 possible rolls of a die.
- **ω** ∈ **Ω** is a **sample point/possible world/atomic event**
- Atomic Event
  - An atomic event is a complete specification of the state of the world.
  - E.g. if the world is only concerned about two Boolean variables A and B. There are four possible atomic events: $A \land B$, $\neg A \land B$, $A \land \neg B$ and $\neg A \land \neg B$ not
- E.g. for the roll of the die, 1, 2 3,.., 6 are all atomic events
- Atomic events are mutually exclusive and collectively exhaustive.
  - $P(\boldsymbol{\omega}_i, \boldsymbol{\omega}_j)=0$ for all $i \neq j$
  - $\Sigma_{(\boldsymbol{\omega} \in \boldsymbol{\Omega})} P(\boldsymbol{\omega}) = 1$

Probability basics (2)

- A **probability space** or **probability model** is a sample space with an assignment $P(\boldsymbol{\omega})$ for every $\boldsymbol{\omega} \in \boldsymbol{\Omega}$ s.t.
  - $0 \leq P(\boldsymbol{\omega}) \leq 1$
  - $\sum_{\boldsymbol{\omega}} P(\boldsymbol{\omega}) = 1$
  - e.g., P(1)=P(2)=P(3)=P(4)=P(5)=P(6)=1/6.
- An **event** A is any subset of $\boldsymbol{\Omega}$
  - $P(A) = \sum_{(\boldsymbol{\omega} \in A)} P(\boldsymbol{\omega})$
  - E.g., P(die roll < 4) = P(1) + P(2) + P(3) = 1/6 + 1/6 + 1/6 = 1/2

Random variables

- A random variable is a variable whose possible values (Domain) are numerical outcomes of a random phenomenon.
  - E.g. X = outcome of a roll of a dice.
  - Domain(X) = {1, 2, 3, 4, 5, 6}
- A Random variable can be boolean, discrete or continuous, depending on its domain.
- ***P*** induces a probability distribution for any r.v. X:
  - e.g., P(X=1) =1/6, P(X=2) =1/6, …, P(X=6) =1/6

Propositions

- Think of a proposition as the event (set of sample points) where the proposition is true
  - Given the random variable, X= outcome of a roll of a dice
- The proposition, X_is_Odd , can be thought of as the event,
  - X_is_Odd = { $\boldsymbol{\omega} \in \boldsymbol{\Omega}$ | X is Odd}
- Proposition = disjunction of atomic events in which it is true. E.g.
  - X_is _Odd ⇔ (X=1) V (X=3) V (X=5)
  - (a V b) ⇔ (¬ a □ b) (a □ ¬ b) (a □ b)

Syntax of Propositions

- **Propositional** or **Boolean** random variables
  - e.g., Cavity (do I have a cavity?)
  - **Cavity =true** is a proposition, also written ***cavity***
- **Discrete** random variables (finite or infinite)
  - e.g., **Weather** is one of **<sunny, rain, cloudy, snow>**
  - **Weather = rain** is a proposition
  - Values must be exhaustive and mutually exclusive
- **Continuous** random variables (bounded or unbounded)
  - e.g., **Temp=21.6**; also allow, e.g., **Temp < 22.0**.
- Arbitrary Boolean combinations of basic propositions

Prior probability

- **Prior** or **unconditional probabilities** of propositions
  - e.g., **P(Cavity =true) = 0.1** and **P(Weather =sunny) = 0.72** correspond to belief prior to arrival of any (new) evidence
- **Probability distribution** gives values for all possible assignments:
  - P(Weather=sunny) = 0.7
  - P(Weather=rain) = 0.2
  - P(Weather=cloudy) = 0.08
  - P(Weather=snow) = 0.02
- **P(Weather) = <0.7, 0.2, 0.08, 0.02>** (normalized, i.e., **sums to 1**)

Prior probability (2)

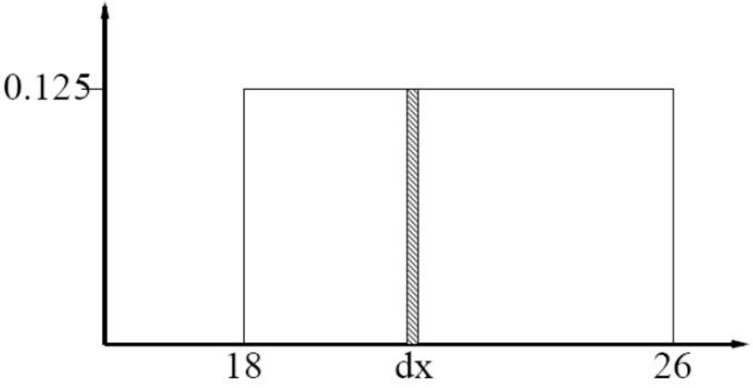
- Joint probability distribution for a set of r.v.s gives the probability of every atomic event on those r.v.s (i.e., every sample point)
- P(Weather, Cavity) is a 4 x 2 matrix of values:

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

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

Probability for continuous variables

- Probability distributions of continuous r.v s are specified by **probability density functions (pdfs)**
  - E.g. Uniform distribution between 18 and 26



Here P is a density; integrates to 1.
$P(X = 20.5) = 0.125$ really means

$$\lim_{dx \to 0} P(20.5 \le X \le 20.5 + dx)/dx = 0.125$$

Joint probability distributions

- A joint distribution is an assignment of probabilities to every possible atomic event

Atomic event	**P**
$Cavity = false \wedge Toothache = false$	0.8
$Cavity = false \wedge Toothache = true$	0.1
$Cavity = true \wedge Toothache = false$	0.05
$Cavity = true \wedge Toothache = true$	0.05

- Suppose we have a joint distribution of n random variables with domain sizes d
  - What is the size of the probability table?
  - Impossible to write out completely for all but the smallest distributions

Notation

- $P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$ refers to a single entry (atomic event) in the joint probability distribution table
  - Shorthand: $P(x_1, x_2, \ldots, x_n)$
- $P(X_1, X_2, \ldots, X_n)$ refers to the entire joint probability distribution table
- $P(A)$ can also refer to the probability of an event
  - E.g., $X_1 = x_1$ is an event

Marginal probability distributions

- From the **joint** distribution **P(X,Y)** we can find the **marginal** distributions **P(X)** and **P(Y)**
- To find P(X = x), sum the probabilities of all atomic events where X = x:

$$P(X = x) = P\big((X = x \wedge Y = y_1) \vee \ldots \vee (X = x \wedge Y = y_n)\big)$$

$$= P\big((x, y_1) \vee \ldots \vee (x, y_n)\big) = \sum_{i=1}^{n} P(x, y_i)$$

- This is called **marginalization** (we are marginalizing out all the variables except X)

Marginal probability distributions

From the **joint** distribution **P(X,Y)** we can find the **marginal** distributions **P(X)** and **P(Y)**

P(Cavity, Toothache)	
$Cavity = false \wedge Toothache = false$	0.8
$Cavity = false \wedge Toothache = true$	0.1
$Cavity = true \wedge Toothache = false$	0.05
$Cavity = true \wedge Toothache = true$	0.05

P(Cavity)	
$Cavity = false$	?
$Cavity = true$	?

P(Toothache)	
$Toothache = false$	?
$Toochache = true$	?

Conditional probability

- Conditional or posterior probabilities
  - e.g., **P(cavity | toothache)** = 0.8
  - i.e., given that toothache is all I know
- If we know more, e.g., there is a rough spot on the tooth, then we have
  - P(cavity | toothache, rough_spot) = 0.9
- New evidence may be irrelevant, allowing simplification, e.g.,
  - P(cavity | toothache, sunny) = P(cavity | toothache) = 0.8
- This kind of inference, sanctioned by domain knowledge, is crucial

Conditional probability (2)

- Definition of co $$P(a / b) = \frac{P(a \wedge b)}{P(b)} \quad \text{if } P(b) > 0$$
- **Product rule** gives an alternative formulation:
  - P(a ∧ b) = P(a | b) P(b) = P(b | a) P(a)
- A general version holds for whole distributions, e.g.,
  - P(Weather, Cavity) = P(Weather | Cavity) P(Cavity)
  - P(A,B) = P(A|B).P(B) = P(B|A).P(A)

Chain rule

Using successive application of the product rule

- Product rule

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

- Chain rule

$$\begin{aligned}
\mathbf{P}(X_1, \ldots, X_n) &= \mathbf{P}(X_n \mid X_{n-1},\ldots,X_1)\ \mathbf{P}(X_{n-1},\ldots,X_1) \\
&= \mathbf{P}(X_n \mid X_{n-1},\ldots,X_1)\ \mathbf{P}(X_{n-1} \mid X_{n-2},\ldots,X_1)\ \mathbf{P}(X_{n-2},\ldots,X_1) \\
&= \mathbf{P}(X_n \mid X_{n-1},\ldots,X_1)\ \mathbf{P}(X_{n-1} \mid X_{n-2},\ldots,X_1)\ \ldots\ \mathbf{P}(X_2/X_1)\ \mathbf{P}(X_1) \\
&= \prod_{i=1}^{n} P(X_i \mid X_{i-1},\ldots,X_1)
\end{aligned}$$

Inference by enumeration

- Start with the **joint probability** distribution:

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	.108	.012	.072	.008
¬ cavity	.016	.064	.144	.576

  - Catch means the dentists steel probe getting caught in the tooth.
- For any **proposition φ**, sum the **atomic events** where it is true:
  - $P(\varphi) = \Sigma_{\omega:\omega \models \varphi} P(\omega)$
  - P(toothache) =?
  - P(cavity V toothache)= ?
  - P(~cavity | toothache) =?

Inference by enumeration (2)

- Start with the joint probability distribution:

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	.108	.012	.072	.008
¬ cavity	.016	.064	.144	.576

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

Inference by enumeration (3)

Start with the joint probability distribution:

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	.108	.012	.072	.008
¬ cavity	.016	.064	.144	.576

For any proposition φ, sum the atomic events where it is true:

$P(\varphi) = \Sigma_{\omega:\omega \models \varphi} P(\omega)$

P(cavity V toothache) = 0.108 + 0.012 +0.072+0.008+ 0.016 + 0.064 = 0.28

Inference by enumeration (4)

- Start with the joint probability distribution:

	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	.108	.012	.072	.008
¬ *cavity*	.016	.064	.144	.576

- Can also compute conditional probabilities:

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

$$= \frac{0.016+0.064}{0.108+0.012+0.016+0.064}$$

$$= 0.4$$

$$P(cavity \mid toothache) = \frac{P(cavity \wedge toothache)}{P(toothache)} \qquad = 0.6$$

Normalization

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	.108	.012	.072	.008
¬ cavity	.016	.064	.144	.576

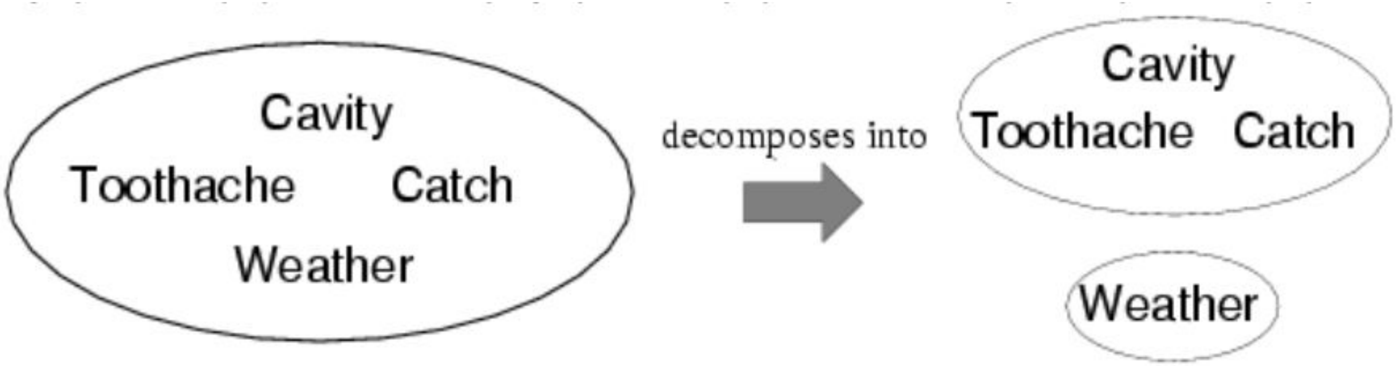
- Denominator can be viewed as a **normalization constant** α
  - P(Cavity | toothache) = αP(Cavity, toothache)
  - = α[P(Cavity, toothache, catch) + P(Cavity, toothache, ~catch)]
  - = α[<0.108, 0.016> + <0.012 ,0.064>]
  - = α<0.12,0.08> = <0.6,0.4>
- General idea: compute distribution on query variable by fixing **evidence variables** and summing over **hidden variables**

Inference by enumeration (5)

- Let **X** be the set of all the variables.
  - Typically, we want 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:
  - $P(Y \mid E = e) = \boldsymbol{\alpha} P(Y, E = e) = \boldsymbol{\alpha\Sigma}_h P(Y, E= e, H = h)$
- The terms in the summation are joint entries because Y, E and H together exhaust the set of random variables
- Obvious problems:
  - Worst-case time complexity O(2n)

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(Toothache, Catch, Cavity, Weather) = P(Toothache, Catch, Cavity) P(Weather)

- 32 entries reduced to 12; for n independent biased coins, $2^n \rightarrow n$
- 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(Toothache, Cavity, Catch) has $2^3 = 8$ independent entries
- If someone has a cavity, the probability that a probe catches in it doesn't depend on whether he has a toothache:
  - P(catch | toothache, cavity) = P(catch | cavity)
- The same independence holds if he hasn’t got a cavity:
  - P(catch | toothache,~cavity) = P(catch | ~cavity)
- ***Catch*** is **conditionally independent** of ***Toothache*** given ***Cavity***:
- P(Catch | Toothache,Cavity) = P(Catch | Cavity)
- Equivalent statements:
  - P(Toothache | Catch, Cavity) = P(Toothache | Cavity)
  - P(Toothache, Catch | Cavity) = P(Toothache | Cavity) P(Catch | Cavity)

Conditional independence (2)

- Write out full joint distribution using chain rule:
- P(Toothache, Catch, Cavity)
  - = P(Toothache | Catch, Cavity) P(Catch, Cavity)
  - = P(Toothache | Catch, Cavity) P(Catch | Cavity) P(Cavity)
  - = P(***Toothache*** | ***Cavity***) P(***Catch*** | ***Cavity***) P(***Cavity***)
- I.e., ***2*** + ***2*** + ***1*** = ***5*** independent numbers
- 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∧b) = P(a | b) P(b) = P(b | a) P(a)

Bayes' rule:

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

or in **distribution form**:

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

Useful for assessing diagnostic probability from causal probability:

$$P(Cause \mid Effect) = \frac{P(Effect \mid Cause)P(Cause)}{P(Effect)}$$

E.g., let M be meningitis, S be stiff neck:

P(m|s) = P(s|m) P(m) / P(s) = 0.8 × 0.0001 / 0.1 = 0.0008

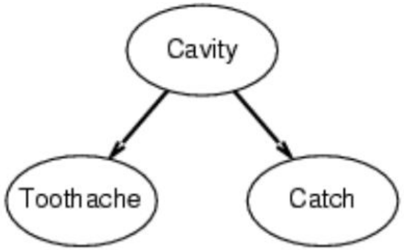
Bayes' rule and conditional independence

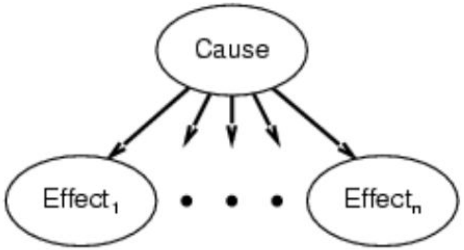
P(Cavity | toothache ∧ catch)

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

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

$P(Cause, Effect_1, \ldots, Effect_n) = P(Cause)\ \pi_i P(Effect_i|Cause)$





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

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
- Independence and conditional independence provide the tools