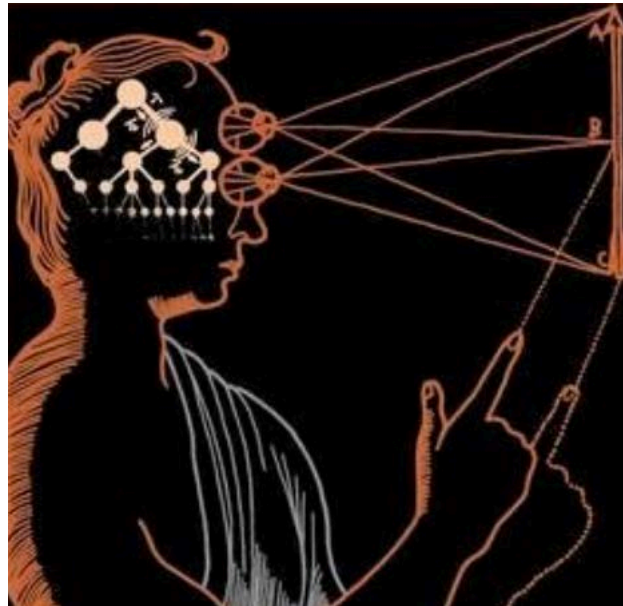


Probabilistic Reasoning

Bang Liu, Jian-Yun Nie

IFT3335: Introduction to Artificial Intelligence



(Adapted from Philipp Koehn, JHU 601.464/664 Artificial Intelligence)

Outline

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

uncertainty

Agents in the real world need to handle uncertainty, whether due to partial observability, nondeterminism, or adversaries. An agent may never know for sure what state it is in now or where it will end up after a sequence of actions.

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
 - uncertainty in action outcomes (flat tire, etc.)
 - immense complexity of modelling and predicting traffic■
- Hence a purely logical approach either
 1. risks falsehood: “ A_{25} will get me there on time”
 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.”



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:

$A_{25} \mapsto_{0.3} \text{AtAirportOnTime}$

$\text{Sprinkler} \mapsto_{0.99} \text{WetGrass}$

$\text{WetGrass} \mapsto_{0.7} \text{Rain}$

Issues: Problems with combination, e.g., *Sprinkler* causes *Rain*? ■

- Probability

Given the available evidence,

A_{25} will get me there on time with probability 0.04

Mahaviracarya (9th C.), Cardano (1565) theory of gambling

- (Fuzzy logic handles **degree of truth** NOT uncertainty e.g.,

WetGrass is true to degree 0.2)

probability

Whereas logical assertions say which possible worlds are strictly ruled out (all those in which the assertion is false), probabilistic assertions talk about how probable the various worlds are.

Probability

- Probabilistic assertions **summarize** effects of
 - laziness**: failure to enumerate exceptions, qualifications, 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., $P(A_{25} | \text{no reported accidents}) = 0.06$
- Might be learned from past experience of similar situations
- Probabilities of propositions change with new evidence:
e.g., $P(A_{25} | \text{no reported accidents, 5 a.m.}) = 0.15$
- Analogous to logical entailment status $KB \models \alpha$, not truth.

Making Decisions under Uncertainty

- Suppose I believe the following:

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

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

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

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

- Which action to choose?
- Depends on my **preferences** for missing flight vs. airport cuisine, etc.
- **Utility theory** is used to represent and infer preferences
- **Decision theory** = utility theory + probability theory
(maximize expected utility)

Probability Basics

- Begin with a set Ω —the sample space
e.g., 6 possible rolls of a die.
 $\omega \in \Omega$ is a sample point/possible world/atomic event■
- A probability space or probability model is a sample space with an assignment $P(\omega)$ for every $\omega \in \Omega$ s.t.
 $0 \leq P(\omega) \leq 1$
 $\sum_{\omega} P(\omega) = 1$
e.g., $P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = 1/6$.■
- An event A is any subset of Ω

$$P(A) = \sum_{\{\omega \in A\}} P(\omega)$$

- E.g., $P(\text{die roll} \leq 3) = P(1) + P(2) + P(3) = 1/6 + 1/6 + 1/6 = 1/2$

Random Variables

- A **random variable** is a function from sample points to some range, e.g., the reals or Booleans
e.g., $Odd(1) = true$.

- P induces a **probability distribution** for any r.v. X :

$$P(X = x_i) = \sum_{\{\omega: X(\omega) = x_i\}} P(\omega)$$

- E.g., $P(Odd = true) = P(1) + P(3) + P(5) = 1/6 + 1/6 + 1/6 = 1/2$

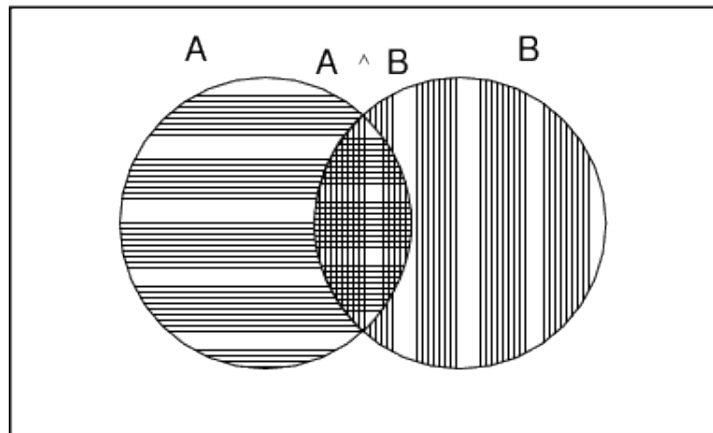
Propositions

- Think of a proposition as the event (set of sample points) where the proposition is true
- Given Boolean random variables A and B :
 - event a = set of sample points where $A(\omega) = true$
 - event $\neg a$ = set of sample points where $A(\omega) = false$
 - event $a \wedge b$ = points where $A(\omega) = true$ and $B(\omega) = true$ ■
- Often in AI applications, the sample points are **defined** by the values of a set of random variables, i.e., the sample space is the Cartesian product of the ranges of the variables■
- With Boolean variables, sample point = propositional logic model
 - e.g., $A = true$, $B = false$, or $a \wedge \neg b$.Proposition = disjunction of atomic events in which it is true
 - e.g., $(a \vee b) \equiv (\neg a \wedge b) \vee (a \wedge \neg b) \vee (a \wedge b)$
 - $\implies P(a \vee b) = P(\neg a \wedge b) + P(a \wedge \neg b) + P(a \wedge b)$

Why use Probability?

- The definitions imply that certain logically related events must have related probabilities
- E.g., $P(a \vee b) = P(a) + P(b) - P(a \wedge b)$

True



Syntax for 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 $\langle \textit{sunny}, \textit{rain}, \textit{cloudy}, \textit{snow} \rangle$
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:
 $\mathbf{P}(Weather) = \langle 0.72, 0.1, 0.08, 0.1 \rangle$ (normalized, i.e., sums to 1)■
- 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)
 $\mathbf{P}(Weather, Cavity) =$ a 4×2 matrix of values:

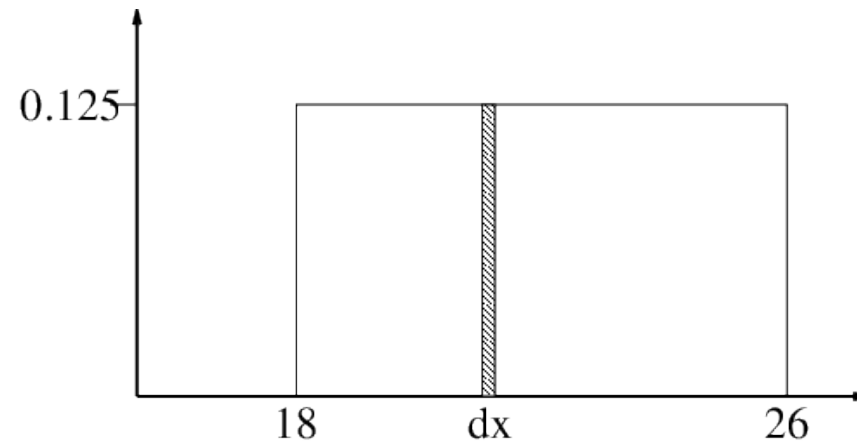
<i>Weather =</i>	<i>sunny</i>	<i>rain</i>	<i>cloudy</i>	<i>snow</i>
<i>Cavity = true</i>	0.144	0.02	0.016	0.02
<i>Cavity = false</i>	0.576	0.08	0.064	0.08

- **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

- Express distribution as a parameterized function of value:

$$P(X = x) = U[18, 26](x) = \text{uniform density between } 18 \text{ and } 26$$



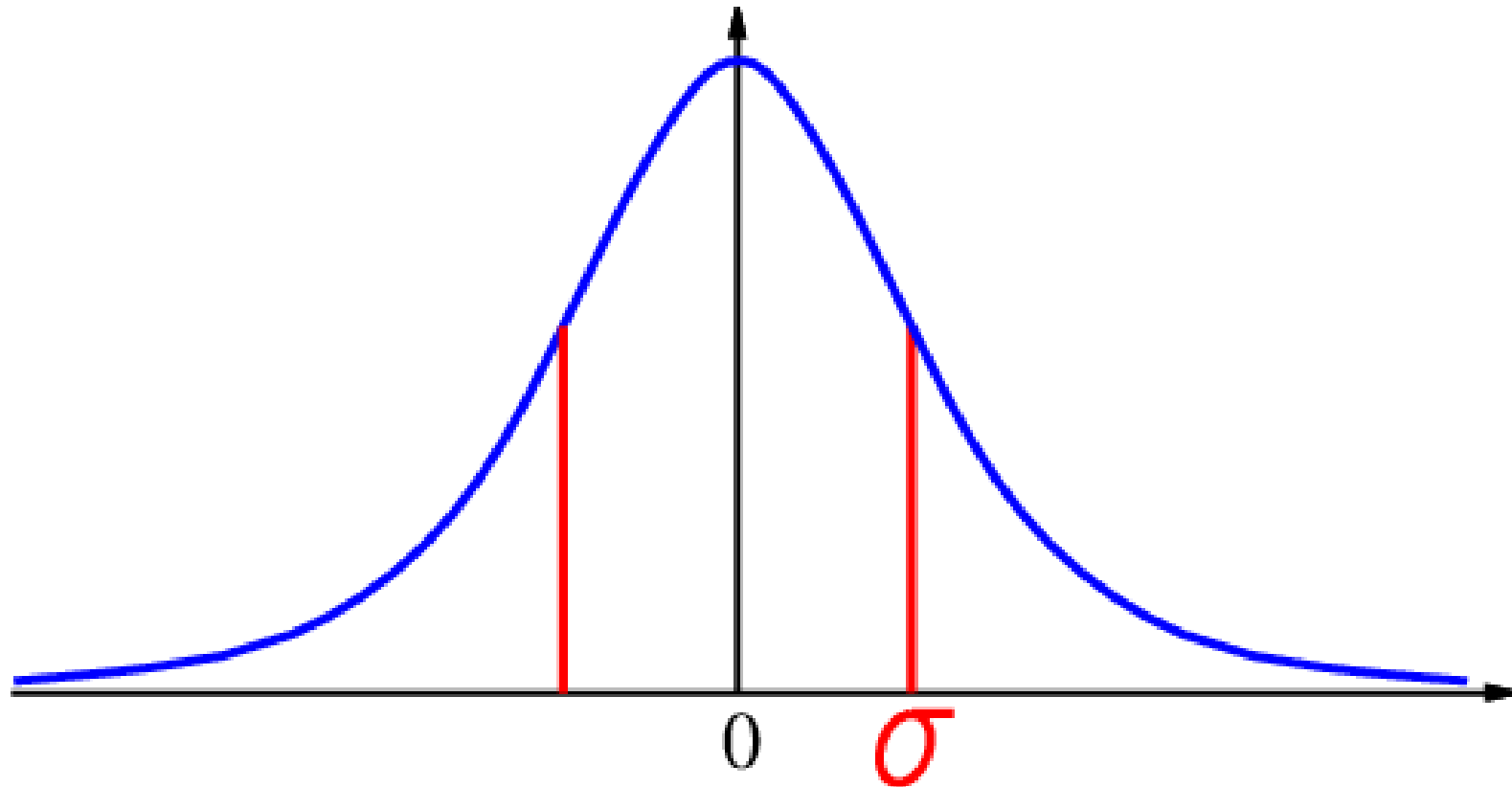
- Here P is a **density**; integrates to 1.

$$P(X = 20.5) = 0.125 \text{ really means}$$

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

Gaussian Density

$$P(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$$



Gaussian distribution is the most important probability distribution in statistics because it fits many natural phenomena like age, height, test-scores, IQ scores, sum of the rolls of two dices and so on.

GN4480100S8

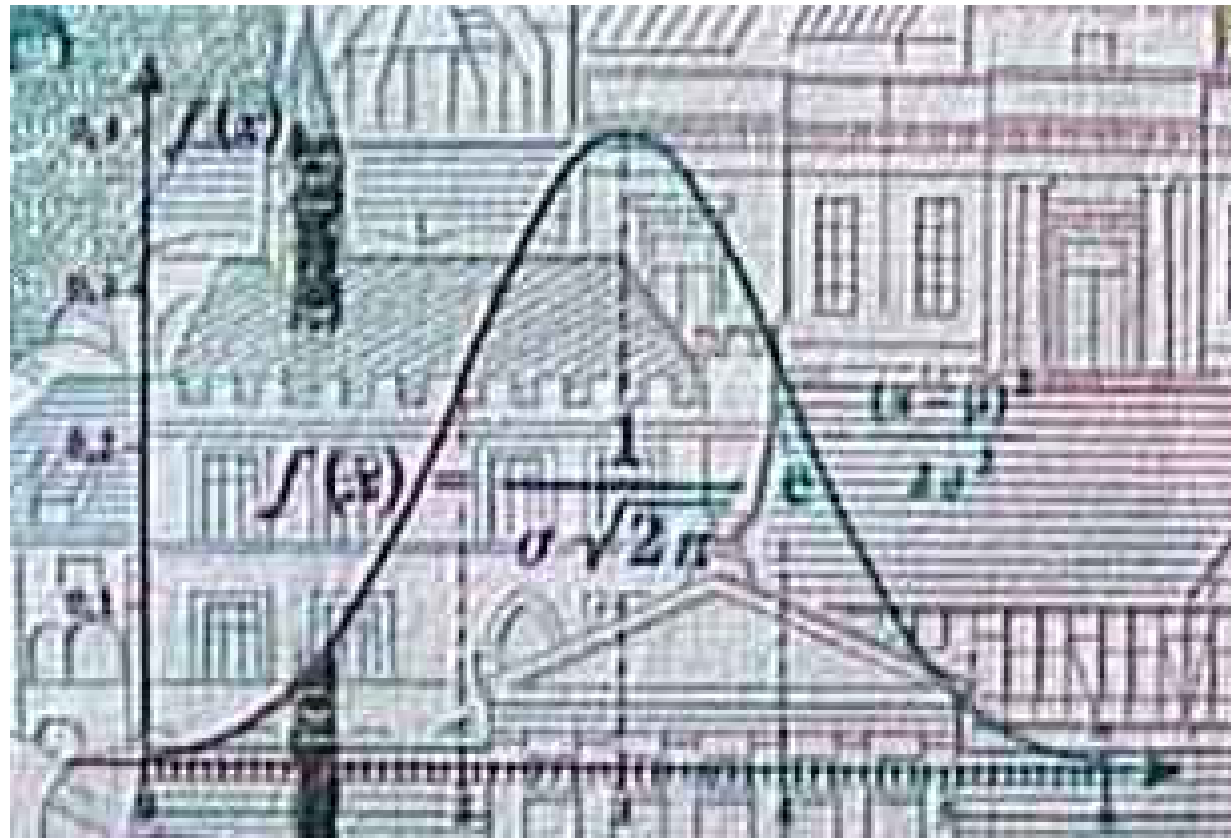
Deutsche Bundesbank

Wolfgang Krauß

Frankfurt am Main
1. September 1999



GN4480100S8



inference

The computation of posterior probabilities for query propositions
given observed evidence.

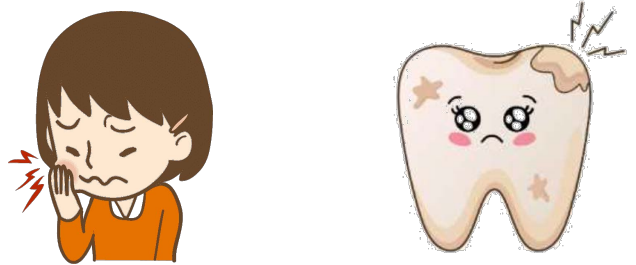
Conditional Probability

- Conditional or posterior probabilities

e.g., $P(\text{cavity}|\text{toothache}) = 0.8$

i.e., **given that** *toothache* **is all I know**

NOT “if *toothache* then 80% chance of *cavity*”



- If we know more, e.g., *cavity* is also given, then we have

$P(\text{cavity}|\text{toothache}, \text{cavity}) = 1$

Note: the less specific belief **remains valid** after more evidence arrives, but is not always **useful**

- New evidence may be irrelevant, allowing simplification, e.g.,

$P(\text{cavity}|\text{toothache}, \text{RavensWin}) = P(\text{cavity}|\text{toothache}) = 0.8$

This kind of inference, sanctioned by domain knowledge, is crucial

Conditional Probability

- Definition of conditional probability:

$$P(a|b) = \frac{P(a \wedge b)}{P(b)} \text{ if } P(b) \neq 0 \blacksquare$$

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

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

- A general version holds for whole distributions, e.g.,

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

(View as a 4×2 set of equations, **not** matrix multiplication) \blacksquare

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

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

Inference by Enumeration

- Start with the joint 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>	.108	.012	.072	.008
\neg <i>cavity</i>	.016	.064	.144	.576

(catch = dentist's steel probe gets caught in cavity)

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

$$P(\phi) = \sum_{\omega:\omega\models\phi} P(\omega)$$

Inference by Enumeration

- Start with the joint 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>	.108	.012	.072	.008
\neg <i>cavity</i>	.016	.064	.144	.576

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

$$P(\phi) = \sum_{\omega: \omega \models \phi} P(\omega)$$

$$P(\text{toothache}) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2$$

Inference by Enumeration

- Start with the joint 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>	.108	.012	.072	.008
\neg <i>cavity</i>	.016	.064	.144	.576

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

$$P(\phi) = \sum_{\omega:\omega\models\phi} P(\omega)$$

$$P(\text{cavity} \vee \text{toothache}) = 0.108 + 0.012 + 0.072 + 0.008 + 0.016 + 0.064 = 0.28$$

Inference by Enumeration

- Start with the joint 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>	.108	.012	.072	.008
\neg <i>cavity</i>	.016	.064	.144	.576

- Can also compute conditional probabilities:

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

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

- Denominator can be viewed as a normalization constant α

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

$$= \alpha [\mathbf{P}(Cavity, toothache, catch) + \mathbf{P}(Cavity, toothache, \neg catch)]$$

$$= \alpha [\langle 0.108, 0.016 \rangle + \langle 0.012, 0.064 \rangle]$$

$$= \alpha \langle 0.12, 0.08 \rangle = \langle 0.6, 0.4 \rangle$$

We can calculate $\mathbf{P}(Cavity|toothache)$ even if we don't know the value of $\mathbf{P}(toothache)$!

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

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

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

- The terms in the summation are joint entries because \mathbf{Y} , \mathbf{E} , and \mathbf{H} together exhaust the set of random variables

- Obvious problems

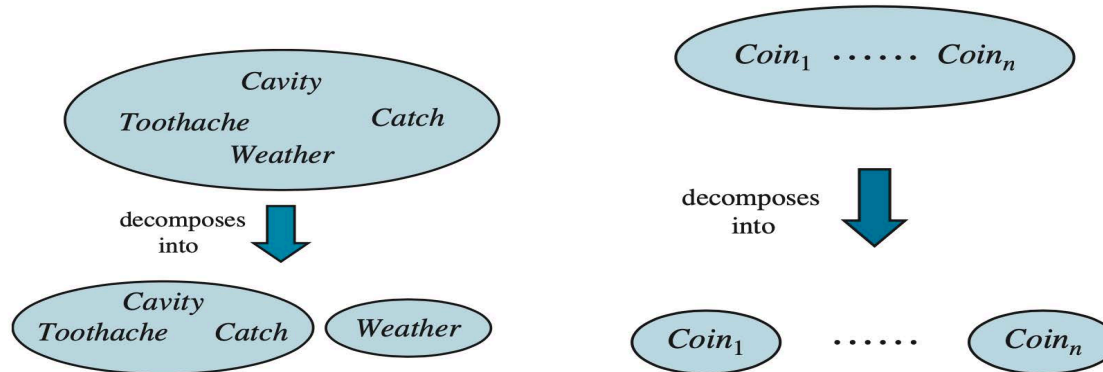
- Worst-case time complexity $O(d^n)$ where d is the largest arity
- Space complexity $O(d^n)$ to store the joint distribution
- How to find the numbers for $O(d^n)$ entries???

independence

Independence assertions can help in reducing the size of the domain representation and the complexity of the inference problem.

Independence

- A and B are independent iff
 $\mathbf{P}(A|B) = \mathbf{P}(A)$ or $\mathbf{P}(B|A) = \mathbf{P}(B)$ or $\mathbf{P}(A, B) = \mathbf{P}(A)\mathbf{P}(B)$

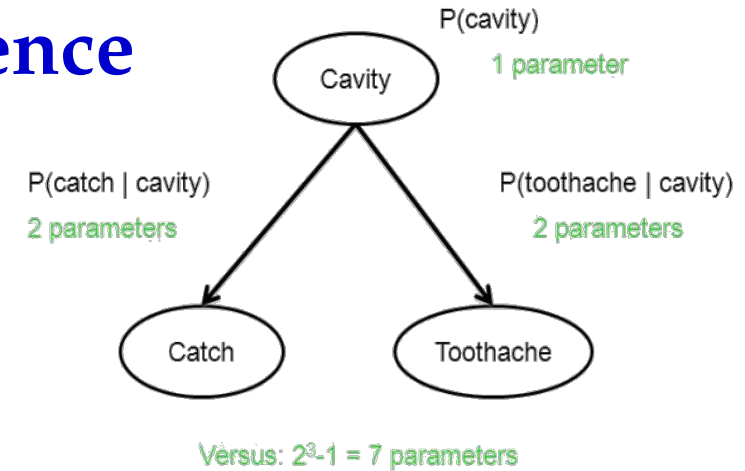


- $\mathbf{P}(\textit{Toothache}, \textit{Catch}, \textit{Cavity}, \textit{Weather})$
 $= \mathbf{P}(\textit{Toothache}, \textit{Catch}, \textit{Cavity})\mathbf{P}(\textit{Weather})$
- 32 entries reduced to 12; for n independent biased coins, $2^n \rightarrow n$
- Absolute independence powerful but rare
- Dentistry is a large field with hundreds of variables, none of which are independent. What to do?

Conditional Independence

- $\mathbf{P}(Toothache, Cavity, Catch)$ 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:
 - (1) $P(catch|toothache, cavity) = P(catch|cavity)$
- The same independence holds if I haven't got a cavity:
 - (2) $P(catch|toothache, \neg cavity) = P(catch|\neg cavity)$
- *Catch* is conditionally independent of *Toothache* given *Cavity*:
$$\mathbf{P}(Catch|Toothache, Cavity) = \mathbf{P}(Catch|Cavity)$$
- Equivalent statements:
$$\mathbf{P}(Toothache|Catch, Cavity) = \mathbf{P}(Toothache|Cavity)$$
$$\mathbf{P}(Toothache, Catch|Cavity) = \mathbf{P}(Toothache|Cavity)\mathbf{P}(Catch|Cavity)$$

Conditional Independence



- Write out full joint distribution using chain rule:

$$\begin{aligned} & \mathbf{P}(\textit{Toothache}, \textit{Catch}, \textit{Cavity}) \\ &= \mathbf{P}(\textit{Toothache} | \textit{Catch}, \textit{Cavity}) \mathbf{P}(\textit{Catch}, \textit{Cavity}) \\ &= \mathbf{P}(\textit{Toothache} | \textit{Catch}, \textit{Cavity}) \mathbf{P}(\textit{Catch} | \textit{Cavity}) \mathbf{P}(\textit{Cavity}) \\ &= \mathbf{P}(\textit{Toothache} | \textit{Cavity}) \mathbf{P}(\textit{Catch} | \textit{Cavity}) \mathbf{P}(\textit{Cavity}) \end{aligned}$$

- I.e., $2 + 2 + 1 = 5$ independent numbers (equations 1 and 2 remove 2)
- 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

In a task such as medical diagnosis, we often have conditional probabilities on causal relationships. The doctor knows $P(\text{symptoms} \mid \text{disease})$ and want to derive a diagnosis, $P(\text{disease} \mid \text{symptoms})$.
Also, diagnostic knowledge is often more fragile than causal knowledge.

Bayes' Rule

- Product rule $P(a \wedge b) = P(a|b)P(b) = P(b|a)P(a)$

$$\implies \text{Bayes' rule } P(a|b) = \frac{P(b|a)P(a)}{P(b)}$$

- Or in distribution form

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

Bayes' Rule

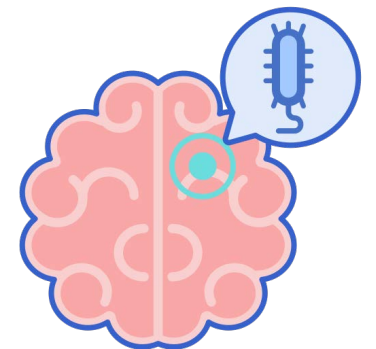
- Useful for assessing **diagnostic** probability from **causal** probability

$$P(\text{Cause}|\text{Effect}) = \frac{P(\text{Effect}|\text{Cause})P(\text{Cause})}{P(\text{Effect})}$$

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

$$P(m|s) = \frac{P(s|m)P(m)}{P(s)} = \frac{0.8 \times 0.0001}{0.1} = 0.0008$$

- Note: posterior probability of meningitis still very small!
(Because the prior probability of stiff necks (from any cause) is much higher than the prior for meningitis.)



Bayes' Rule and Conditional Independence

- Example of a naive Bayes model

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

- Generally:

$$\mathbf{P}(Cause, Effect_1, \dots, Effect_n) = \mathbf{P}(Cause) \prod_i \mathbf{P}(Effect_i|Cause)$$



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

wampus world

We can combine the ideas in this chapter to solve probabilistic reasoning problems in the wampus world. Uncertainty arises in the wampus world because the agent's sensors give only partial information about the world.

Wumpus World

1,4	2,4	3,4	4,4
1,3	2,3	3,3	4,3
1,2 B OK	2,2	3,2	4,2
1,1 OK	2,1 B OK	3,1	4,1

- $P_{ij} = true$ iff $[i, j]$ contains a pit
- $B_{ij} = true$ iff $[i, j]$ is breezy
Include only $B_{1,1}, B_{1,2}, B_{2,1}$ in the probability model

Specifying the Probability Model

- The full joint distribution is $\mathbf{P}(P_{1,1}, \dots, P_{4,4}, B_{1,1}, B_{1,2}, B_{2,1})$ ■
- Apply product rule: $\mathbf{P}(B_{1,1}, B_{1,2}, B_{2,1} | P_{1,1}, \dots, P_{4,4}) \mathbf{P}(P_{1,1}, \dots, P_{4,4})$

This gives us: $P(\textit{Effect} | \textit{Cause})$ ■

- First term: 1 if pits are adjacent to breezes, 0 otherwise■
- Second term: pits are placed randomly, probability 0.2 per square:

$$\mathbf{P}(P_{1,1}, \dots, P_{4,4}) = \prod_{i,j=1,1}^{4,4} \mathbf{P}(P_{i,j}) = 0.2^n \times 0.8^{16-n}$$

for n pits.

Observations and Query

- We know the following facts:

$$b = \neg b_{1,1} \wedge b_{1,2} \wedge b_{2,1}$$

$$known = \neg p_{1,1} \wedge \neg p_{1,2} \wedge \neg p_{2,1}$$

- Query is $\mathbf{P}(P_{1,3}|known, b)$
- Define $Unknown = P_{ij}$ s other than $P_{1,3}$ and $Known$
- For inference by enumeration, we have

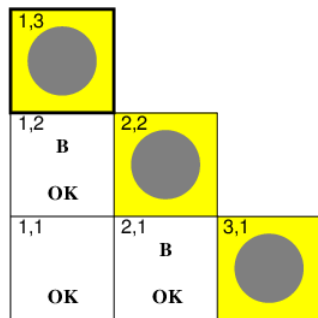
$$\mathbf{P}(P_{1,3}|known, b) = \alpha \sum_{unknown} \mathbf{P}(P_{1,3}, unknown, known, b)$$

- Grows exponentially with number of squares!

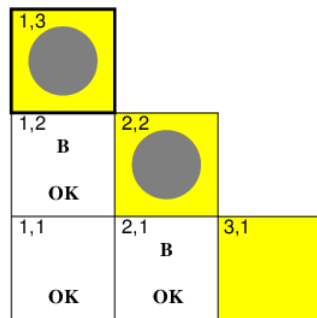
Using Conditional Independence

$$\begin{aligned}
 \mathbf{P}(P_{1,3}|known, b) &= \alpha \sum_{unknown} \mathbf{P}(P_{1,3}, unknown, known, b) \\
 &\quad \text{(use product rule)} \\
 &= \alpha \sum_{unknown} \mathbf{P}(b|P_{1,3}, known, unknown) \mathbf{P}(P_{1,3}, known, unknown) \blacksquare \\
 &\quad \text{(unknown = fringe + other)} \\
 &= \alpha \sum_{fringe} \sum_{other} \mathbf{P}(b|known, P_{1,3}, fringe, other) \mathbf{P}(P_{1,3}, known, fringe, other) \blacksquare \\
 &\quad \text{(b is independent of other given known)} \\
 &= \alpha \sum_{fringe} \sum_{other} \mathbf{P}(b|known, P_{1,3}, fringe) \mathbf{P}(P_{1,3}, known, fringe, other) \blacksquare \\
 &\quad \text{(move the summation inward)} \\
 &= \alpha \sum_{fringe} \mathbf{P}(b|known, P_{1,3}, fringe) \sum_{other} \mathbf{P}(P_{1,3}, known, fringe, other) \blacksquare \\
 &\quad \text{(right term can be factored by independence)} \\
 &= \alpha \sum_{fringe} \mathbf{P}(b|known, P_{1,3}, fringe) \sum_{other} \mathbf{P}(P_{1,3}) P(known) P(fringe) P(other) \blacksquare \\
 &\quad \text{(reorder terms)} \\
 &= \alpha P(known) \mathbf{P}(P_{1,3}) \sum_{fringe} \mathbf{P}(b|known, P_{1,3}, fringe) P(fringe) \sum_{other} P(other) \blacksquare \\
 &\quad \text{(folds P(known) into a')} \\
 &= \alpha' \mathbf{P}(P_{1,3}) \sum_{fringe} \mathbf{P}(b|known, P_{1,3}, fringe) P(fringe) \quad (\sum P(other) = 1)
 \end{aligned}$$

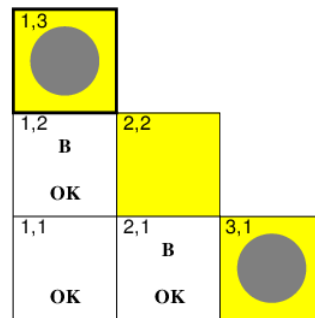
Using Conditional Independence



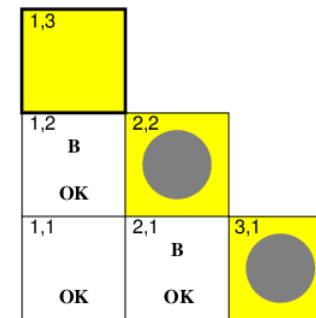
$$0.2 \times 0.2 = 0.04$$



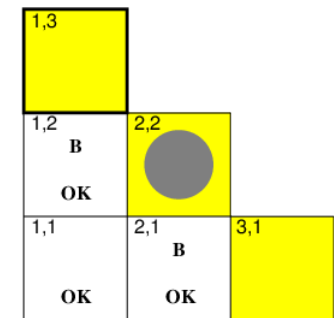
$$0.2 \times 0.8 = 0.16$$



$$0.8 \times 0.2 = 0.16$$



$$0.2 \times 0.2 = 0.04$$



$$0.2 \times 0.8 = 0.16$$

$$\begin{aligned} \mathbf{P}(P_{1,3} | \textit{known}, b) &= \alpha' \langle 0.2(0.04 + 0.16 + 0.16), 0.8(0.04 + 0.16) \rangle \\ &\approx \langle 0.31, 0.69 \rangle \end{aligned}$$

$$\mathbf{P}(P_{2,2} | \textit{known}, b) \approx \langle 0.86, 0.14 \rangle$$

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