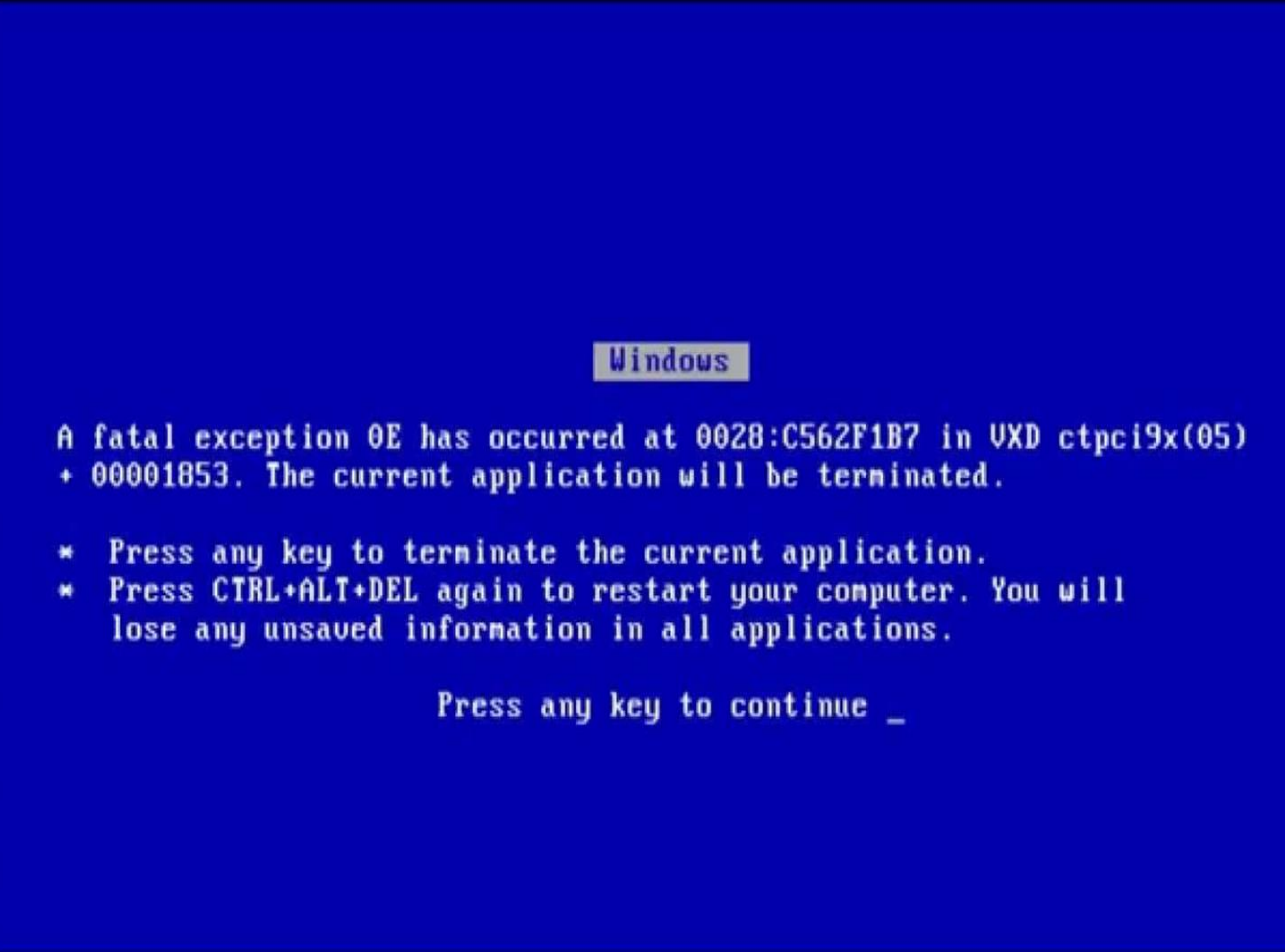
Artificial Intelligence
Chapter 13: Quantifying Uncertainty


Windows
A fatal exception 0E has occurred at 0028:C562F1B7 in VXD ctpci9x(05) + 00001853. The current application will be terminated.
* Press any key to terminate the current application.
* Press CTRL+ALT+DEL again to restart your computer. You will lose any unsaved information in all applications.
Press any key to continue _

A fatal exceptio ... in VXD ctpci9x(05)
+ 00001853. The c ... ted.

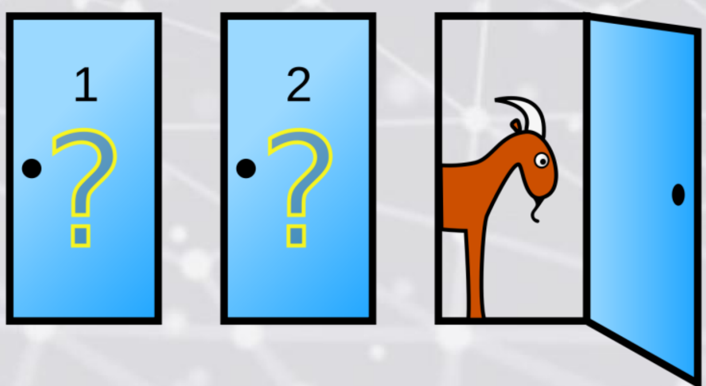
* Press any key ... on.
* Press CTRL+AL ... . You will
lose any unsav ... .

Digression: The Monty Hall Problem

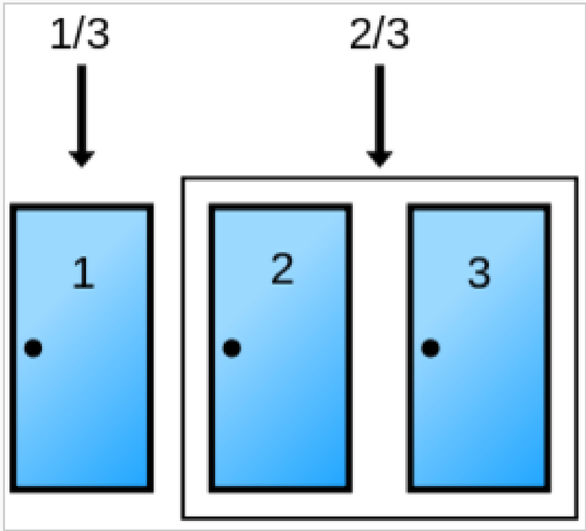
- Suppose you're on a game show, and you're given the choice of three doors:

Behind one door is a car; behind the others, goats.

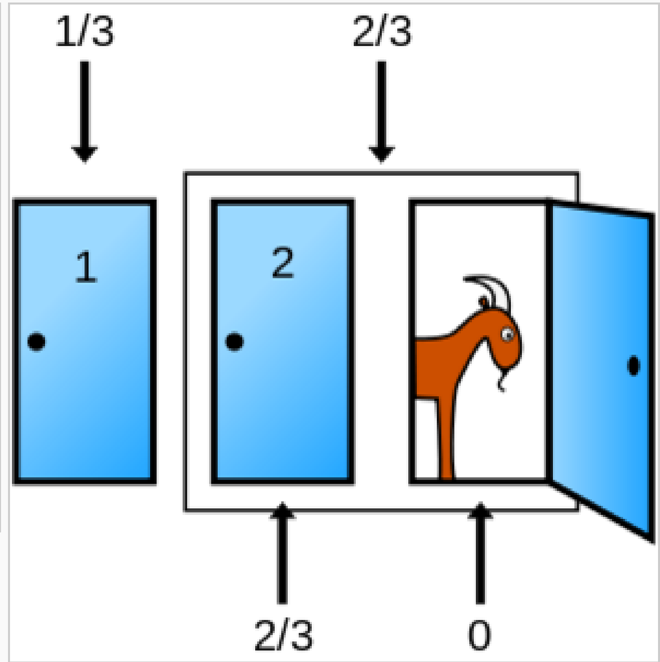
You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat. He then says to you, "Do you want to pick door No. 2?" Is it to your advantage to switch your choice?



Digression: The Monty Hall Problem

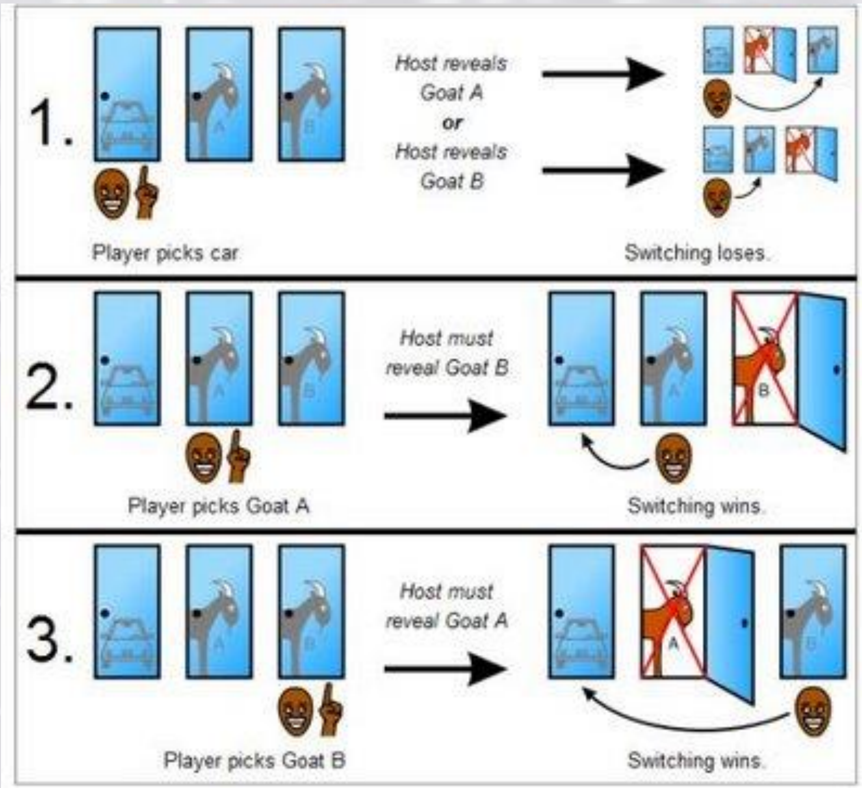


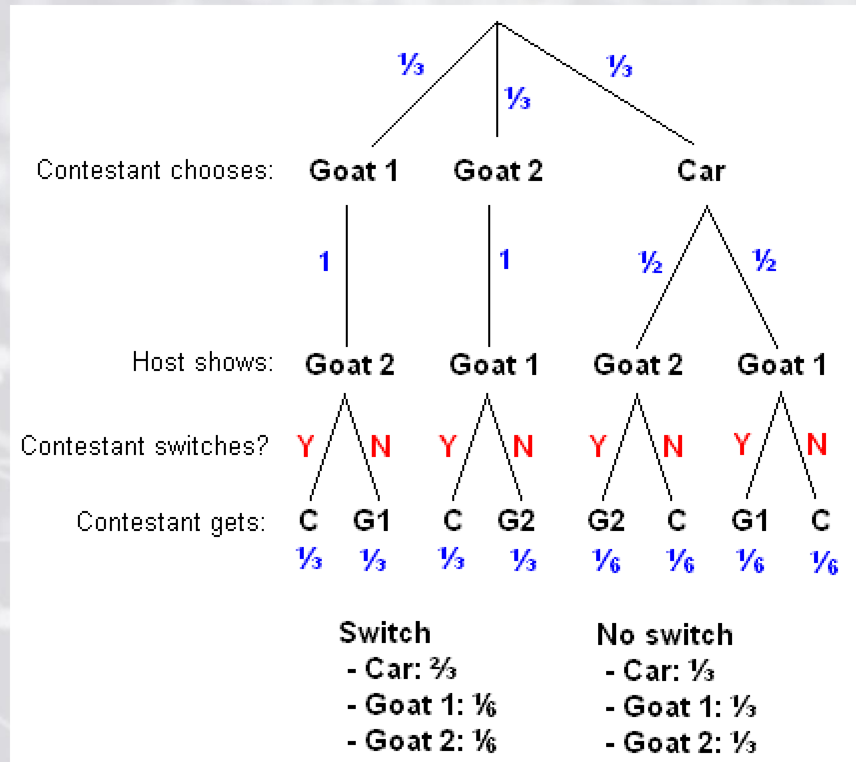
Car has a 1/3 chance of being behind the player's pick and a 2/3 chance of being behind one of the other two doors.



The host opens a door, the odds for the two sets don't change but the odds move to 0 for the open door and 2/3 closed door.

Digression: The Monty Hall Problem





Uncertainty

- Agents need to handle uncertainty, whether due to partial observability, non-determinism, or a combination of the two.

- In Chapter 4, we encountered problem-solving agents designed to handle uncertainty by monitoring a **belief-state** – a representation of the set of all possible world states in which the agent might find itself (e.g. AND-OR graphs).

- The agent generated a **contingency plan** that handles every possible eventuality that its sensors report during execution.

Uncertainty

- Despite its many virtues, however, this approach has many **significant drawbacks**:

(*) With partial information, an agent must consider *every* possible eventuality, no matter how unlikely. This leads to impossibly large and complex belief-state representations.

(*) A correct contingency plan that handles every possible outcome can grow arbitrarily large and must consider arbitrarily unlikely contingencies.

(*) Sometimes there is, in fact, no plan that is guaranteed to achieve a stated goal – yet the agent must act. It must have some way to compare the merits of plans that are not guaranteed.

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 …)

Uncertainty

- Consider a trivial example of uncertain reasoning for medical diagnosis.

(*) Toothache => Cavity (this is faulty)

Me amend it:

(*) Tootache => Cavity V Gum Problem V Abscess…

Problem is that we would need to add an almost unlimited list of possible symptoms.

- We could instead attempt to turn the rule into a *causal rule*.

(*) Cavity => Tootache (this is also incorrect; not all cavities cause pain).

Uncertainty

- The only way to fix the rule, it seems, is to make it logically exhaustive! (i.e. augment the left-hand side with all the qualifications required for a cavity to cause a toothache).

- This approach though naturally fails for at least (3) reasons:

(1) **Laziness:** far too much work is required to compile the entire list.

(2) **Theoretical Ignorance:** Medical science is theoretically incomplete.

(3) **Practical Ignorance:** Even if we knew all the rules, we might be uncertain about a particular patient, because not all of the necessary tests have been run.

Uncertainty

- Typically, an agent’s knowledge can at best provide only a **degree of belief**.

- Our main tool for dealing with degrees of belief is **probability theory**.

- Probability provides a way of <u>summarizing the uncertainty</u> that comes from our *laziness* and *ignorance*.

Uncertainty and Rational Decisions

- So how best can an agent make rational decisions in the face of uncertainty?
- To make choices, the agent must first have **preferences** between possible **outcomes** of the various plans.
- An outcome is a completely specified state, including such factors as whether the agent arrives on time (e.g. the "airport problem").
- We use **utility theory** to represent reason with preferences. Utility theory asserts that every state has a degree of *usefulness*, or utility, to an agent and that the agent will prefer states with higher utility.

Uncertainty and Rational Decisions

- Preferences, as expressed by utilities, are combined with probabilities in the general theory of rational decisions called **decision theory**:

Decision Theory = Probability Theory + Utility Theory

- Fundamental idea: *an agent is **rational** iff it chooses the action that yields the highest expected utility, averaged over all possible outcomes of the action*. (The principle of maximum expected utility (**MEU**)).

- Note that this is none other than a computation of **expected value**.

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., $P(A_{25} \mid \text{no reported accidents}) = 0.06$

These are not assertions about the world

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:

$P(A_{25}$ gets me there on time $| \ldots) = 0.04$
$P(A_{90}$ gets me there on time $| \ldots) = 0.70$
$P(A_{120}$ gets me there on time $| \ldots) = 0.95$
$P(A_{1440}$ gets me there on time $| \ldots) = 0.9999$

- 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

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 <*sunny,rainy,cloudy,snow*>
- Domain values must be exhaustive and mutually exclusive
- Elementary proposition constructed by assignment of a value to a
- random variable: e.g., *Weather* = *sunny*, *Cavity* = *false*
- (abbreviated as ¬*cavity*)
- Complex propositions formed from elementary propositions and standard logical connectives e.g., *Weather* = *sunny* ∨ *Cavity* = *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

Axioms of probability

- The set of all possible "worlds" is the **sample space (omega)**. The possible worlds are **mutually exclusive** and **exhaustive**.
- A fully specified probability model associates a numerical probability *P(ω)* with each possible world (we assume discrete, countable worlds).

$$0 \le P(\omega) \le 1 \;\; \textit{for every } \omega, \;\; \textit{and} \;\; \sum_{\omega \in \Omega} P(w) = 1$$

Axioms of probability

- Probabilistic assertions are usually about **sets** instead of particular possible worlds.
- These sets are commonly referred to as **events**.

- In AI, the sets are described by propositions in a formal language. The probability associated with a proposition is defined to be the sum of probabilities of the worlds in which it holds:

$$\text{For any proposition } \phi,\ P(\phi) = \sum_{\omega \in \phi} P(\omega)$$

Axioms of probability

- Probabilistic assertions are usually about **sets** instead of particular possible worlds.
- These sets are commonly referred to as **events**.

- In AI, the sets are described by propositions in a formal language. The probability associated with a proposition is defined to be the sum of probabilities of the worlds in which it holds:

$$\text{For any proposition } \phi,\ P(\phi) = \sum_{\omega \in \phi} P(\omega)$$

Axioms of probability

- The basic axioms of probability imply certain relationships among the degrees of belief that can be accorded to logically-related propositions. Example:

$$
\begin{aligned}
P(\neg a) &= \sum_{\omega \in \neg a} P(\omega) \\
&= \sum_{\omega \in \neg a} P(\omega) + \sum_{\omega \in a} P(\omega) - \sum_{\omega \in a} P(\omega) \\
&= \sum_{\omega \in \Omega} P(\omega) - \sum_{\omega \in a} P(\omega) \\
&= 1 - P(a)
\end{aligned}
$$

Axioms of probability

- **Inclusion-Exclusion** (probability of a *disjunction*):

$$P(a \vee b) = P(a) + P(b) - P(a \wedge b)$$

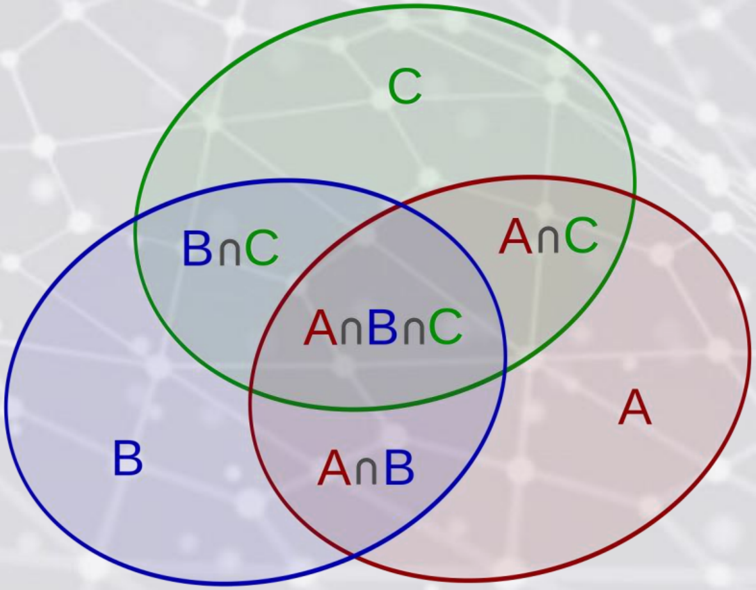
- Now derive the general formula for three or more sets…

Axioms of probability

- **Inclusion-Exclusion** (probability of a *disjunction*):

$$P(a \vee b) = P(a) + P(b) - P(a \wedge b)$$

- Now derive the general formula for three or more sets…



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*) = <0.72,0.1,0.08,0.1> (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(*Weather*,*Cavity*) = a 4 × 2 matrix of values:

Weather =	sunny	rainy	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

Conditional probability

- Conditional or posterior probabilities
  e.g., P(*cavity* | *toothache*) = 0.8

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

- (Notation for conditional distributions:
  P(*Cavity* | *Toothache*) = 2-element vector of 2-element vectors)

- If we know more, e.g., *cavity* is also given, then we have
  P(*cavity* | *toothache,cavity*) = 1

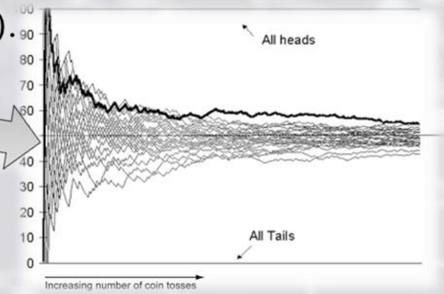
- New evidence may be irrelevant, allowing simplification, e.g.,
  P(*cavity* | *toothache, sunny*) = P(*cavity* | *toothache*) = 0.8
- This kind of inference, sanctioned by domain knowledge, is crucial

Conditional probability

- Definition of conditional probability:
- $P(a \mid b) = P(a \wedge b) / P(b)$ if $P(b) > 0$
- Product rule gives an alternative formulation:
- $P(a \wedge b) = P(a \mid b) P(b) = P(b \mid a) P(a)$
- A general version holds for whole distributions, e.g.,
- $\mathbf{P}(Weather, Cavity) = \mathbf{P}(Weather \mid Cavity)\, \mathbf{P}(Cavity)$
- (View as a set of 4 × 2 equations, not matrix mult.)
-
- Chain rule is derived by successive application of product rule:
- $\mathbf{P}(X_1, \ldots, X_n) = \mathbf{P}(X_1,\ldots,X_{n-1})\, \mathbf{P}(X_n \mid X_1,\ldots,X_{n-1})$
  $= \mathbf{P}(X_1,\ldots,X_{n-2})\, \mathbf{P}(X_{n-1} \mid X_1,\ldots,X_{n-2})\, \mathbf{P}(X_n \mid X_1,\ldots,X_{n-1})$
  $= \ldots$
  $= \pi_{i=1}$^n $\mathbf{P}(X_i \mid X_1, \ldots, X_{i-1})$

Bayesian and Frequentist Probability

- (2) General paradigms for statistics and statistical inference: *frequentist* vs. *Bayesian.*
- Frequentists: Parameters are fixed; there is a (Platonic) model; parameters remain constant.
- Bayesians: Data are fixed; data are observed from realized sample; we encode prior beliefs; parameters are described probabilistically.
- Frequentists commonly use the *MLE* (**maximum likelihood estimate**) as a cogent *point estimate* of the model parameters of a probability distribution: $\hat{\theta}_{MLE} = \underset{\theta}{\operatorname{argmax}} L(D|\theta)$.
- Using the *Law of Large Numbers (LLN),* $\lim_{n\to\infty} P(|\bar{X}_n - \mu| > \varepsilon) = 0$, one can consequently show that: $\hat{\theta}_{MLE} \xrightarrow{P} \theta$.



Potential issues with frequentist approach: philosophical reliance on long-term ‘frequencies’, *the problem of induction* (Hume) and the black swan paradox, as well as the presence of limited exact solutions for a small class of settings.

Bayesian and Frequentist Probability

In the Bayesian framework, conversely, probability is regarded as a measure of uncertainty pertaining to the practitioner's knowledge about a particular phenomenon.
The prior belief of the experimenter is not ignored but rather encoded in the process of calculating probability.
As the Bayesian gathers new information from experiments, this information is used, in conjunction with prior beliefs, to update the measure of certainty related to a specific outcome.
These ideas are summarized elegantly in the familiar *Bayes' Theorem*:

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

Where H here connotes '*hypothesis*' and D connotes '*data*'; the leftmost probability is referred to as the *posterior* (of the hypothesis), and the numerator factors are called the *likelihood* (of the data) and the *prior* (on the hypothesis), respectively; the denominator expression is referred to as the *marginal likelihood*.

Typically, the point estimate for a parameter used in Bayesian statistics is the *mode* of the *posterior distribution*, known as the **maximum a posterior** (MAP) estimate, which is given as:

$$\hat{\theta}_{MAP} = \underset{\theta}{\text{argmax}} P(D|\theta)P(\theta)$$

Practice Problems

(1) Derive **Inclusion-Exclusion** from Equations (13.1) and (13.2) in the text.

$$P(a \vee b) = P(a) + P(b) - P(a \wedge b)$$

$$0 \leq P(\omega) \leq 1 \ \ \textit{for every } \omega, \ \textit{and} \sum_{\omega \in \Omega} P(w) = 1 \qquad (13.1)$$

$$\textit{For any proposition } \phi, \ P(\phi) = \sum_{\omega \in \phi} P(\omega) \qquad (13.2)$$

(2) Consider the set of all possible five-card poker hands dealt fairly (i.e. randomly) from a single, standard deck.

(i) How many atomic events are there in the joint probability distribution?

(ii) What is the probability of each atomic event?

(iii) What is the probability of being dealt a royal flush?

(iv) Four of a kind?

(v) Given that my first two cards are aces, what is the probability that my total hand consists of four aces?

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

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

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

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

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

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

- P(*toothache*) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2

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

- Can also compute conditional probabilities:
-

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

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

$$= 0.4$$

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*) = α, **P**(*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

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**

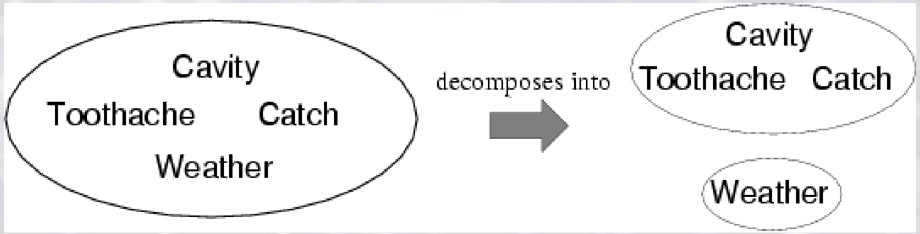
Then the required summation of joint entries is done by summing out the hidden variables:

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

- 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?

Independence

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



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

- 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

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

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

Conditional independence

- Write out full joint distribution using chain rule:
-

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

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

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

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

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

Practice Problems II

(1) Show that the (3) forms of "absolute" independence are equivalent.

$$\mathbf{P}(A|B) = \mathbf{P}(A) \quad \text{or } \mathbf{P}(B|A) = \mathbf{P}(B) \quad \text{or } \mathbf{P}(\mathrm{A, B}) = \mathbf{P}(A)\,\mathbf{P}(B)$$

(2) Suppose that X, Y are independent random variables; let Z be a function of X and Y. Must X and Y be conditionally independent, given Z? Explain.

(3) Suppose you are given a bag containing n unbiased coins. You are told that n – 1 of these coins are "normal", with heads on one side and tails on the other, whereas one coin is a fake, with heads on both sides.

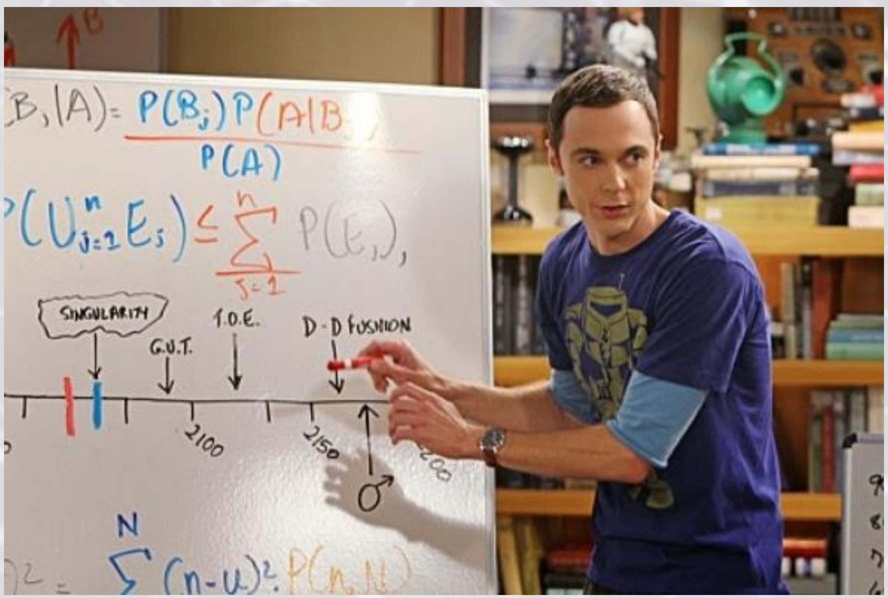
Consider the scenario in which you reach into the bag, pick out a coin at random, flip it, and get a head. What is the (conditional) probability that the coin you chose is fake?

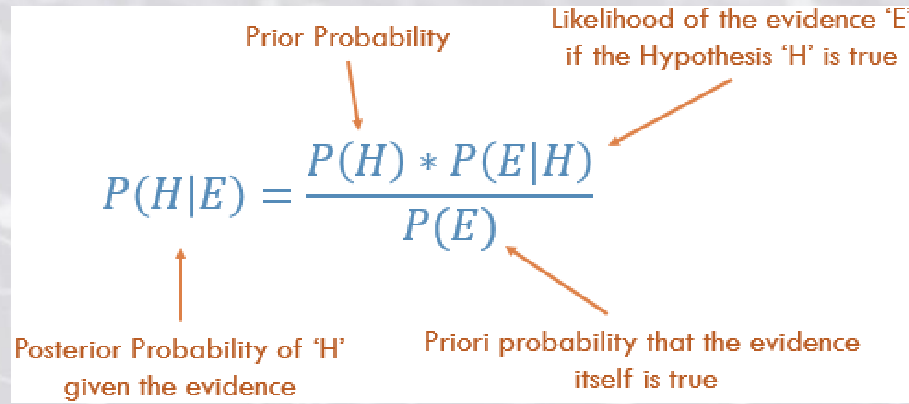
Bayes' Rule

- Product rule $P(a \land b) = P(a \mid b)\, P(b) = P(b \mid a)\, P(a)$
-

$\Rightarrow$ Bayes' rule: $P(a \mid b) = P(b \mid a)\, P(a) / P(b)$

Derive Bayes' Rule…





Bayes' Rule

- In distribution form:

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

- Useful for assessing diagnostic probability from causal probability:

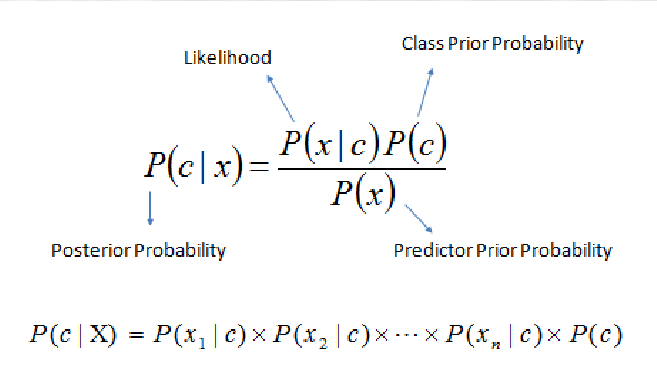
  - P(Cause|Effect) = P(Effect|Cause) P(Cause) / P(Effect)
  - E.g., let M be meningitis, S be stiff neck:
  - $P(m|s) = P(s|m)\,P(m) / P(s) = 0.8 \times 0.0001 / 0.1 = 0.0008$
  - Note: posterior probability of meningitis still very small!

Bayes' Rule and conditional independence

$\mathbf{P}(\mathit{Cavity} \mid \mathit{toothache} \wedge \mathit{catch})$

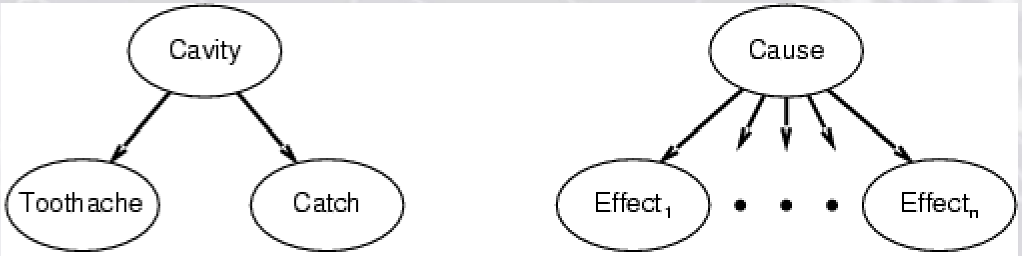
$= \alpha \mathbf{P}(\mathit{toothache} \wedge \mathit{catch} \mid \mathit{Cavity})\ \mathbf{P}(\mathit{Cavity})$

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



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

$$\mathbf{P}(\text{Cause}, \text{Effect}_1, \ldots, \text{Effect}_n) = \mathbf{P}(\text{Cause})\ \pi_i \mathbf{P}(\text{Effect}_i \mid \text{Cause})$$



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

Summary

- Uncertainty arises because of both laziness and ignorance. It is **inescapable** in complex, nondeterministic, or partially observable environments.

- Probability is a rigorous formalism for uncertain knowledge. Probabilities summarize the agent's believes relative to the evidence.

- **Decision Theory** combines the agent's beliefs and desires, defining the best action as the one that maximizes expected utility.

- Basic probability statements include **priors probabilities** and **conditional probabilities**. Joint probabilities distributions specifiy a probability of every atomic event.

Summary

- **Absolute independence** between subsets of random variables allows the full joint distribution to be factored into smaller joint distributions, greatly reducing its complexity. Absolute independence seldom occurs in practice.

- **Bayes' Rule** allows unknown probabilities to be computer from known conditional probabilities, usually in the casual direction.

- **Conditional independence** brought about by direct causal relationships in the domain might allow the full joint distribution to be factored into smaller, conditional distributions.

- The **naïve Bayes** model assumes the conditional independence of all effect variables, given a single cause variable, and grows linearly with the number of effects.