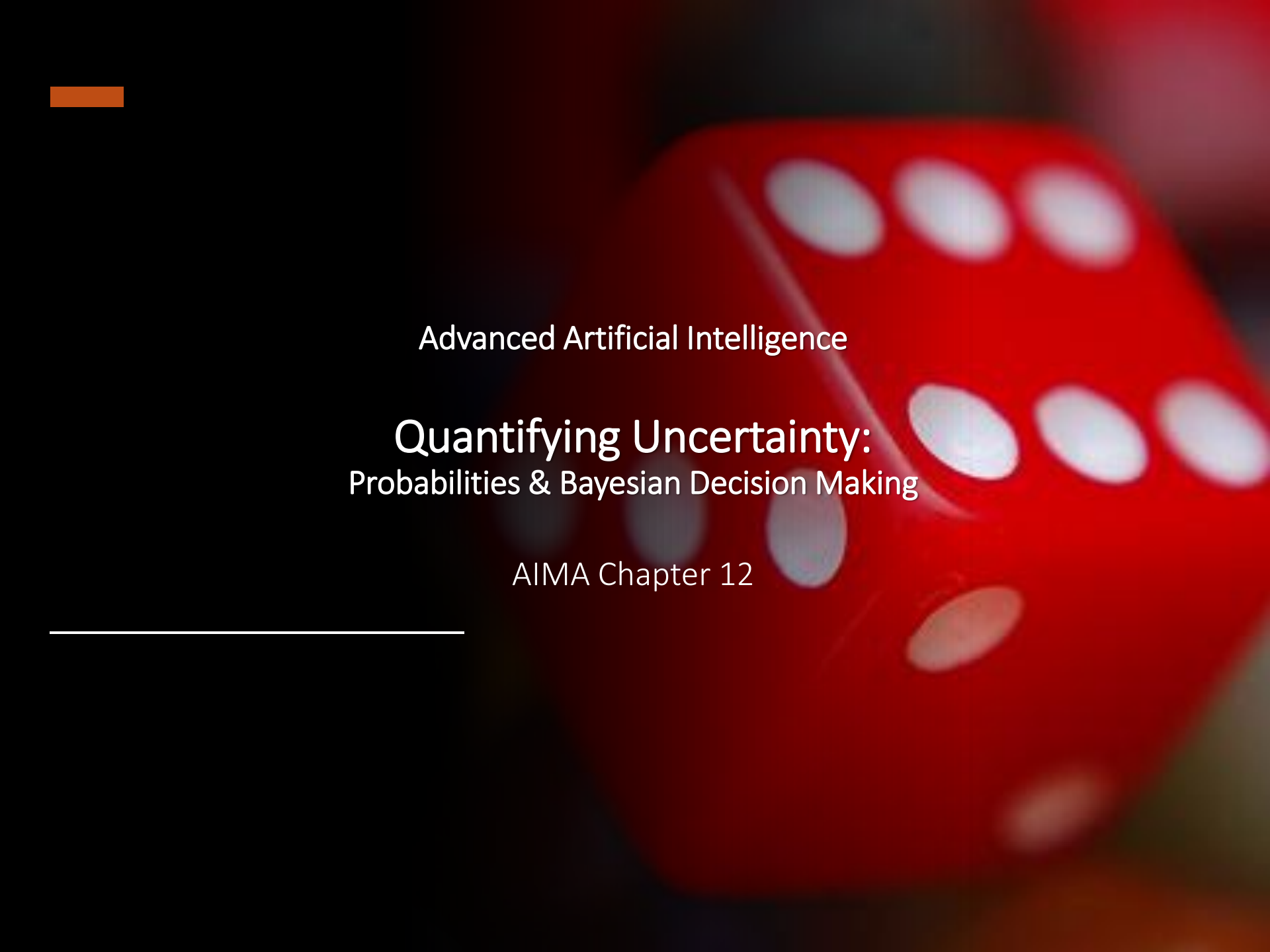


Advanced Artificial Intelligence

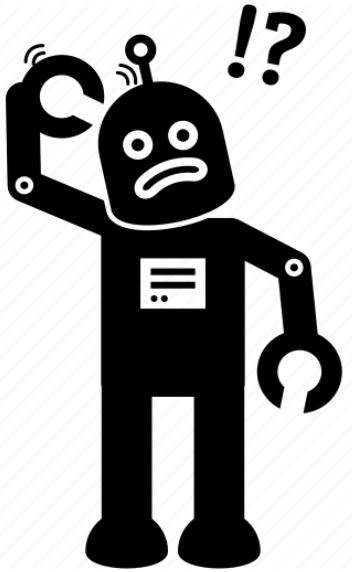
Quantifying Uncertainty:
Probabilities & Bayesian Decision Making

AIMA Chapter 12

Example: Catching a Flight with a Logical Agent

Let action A_t = leave for airport t minutes before flight

Question: What action A_t get me there on time?



Problems:

- Partial observability (road state, other drivers' plans, etc.)
- Noisy sensors (traffic reports)
- Uncertainty in action outcomes (flat tire, etc.)
- Complexity of modeling and predicting traffic

Logic leads to the following conclusions:

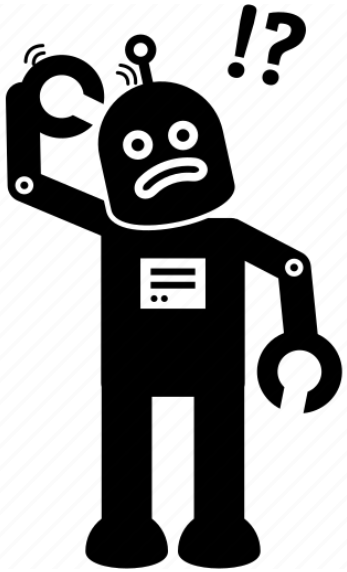
- A_{25} will get me there on time if there is no accident on the bridge and it doesn't rain and my tires remain intact, etc., etc.
- A_{Inf} guarantees to get there in time, but who lives forever?

Logic often creates conclusions that are too weak for effective decision-making.
Uncertainty is a problem for logical agents!

Example: Catching a Flight with Belief States

Let action A_t = leave for airport t minutes before flight

Question: What action A_t get me there on time?



Belief states

- Are used to deal with uncertainty in the environment.
- Are the set of states the agent believes it could be in.
- Often are the result of nondeterministic actions:

$$\text{Results}(\text{at home}, A_t) = \{\text{on time}, \text{missed flight}\}$$

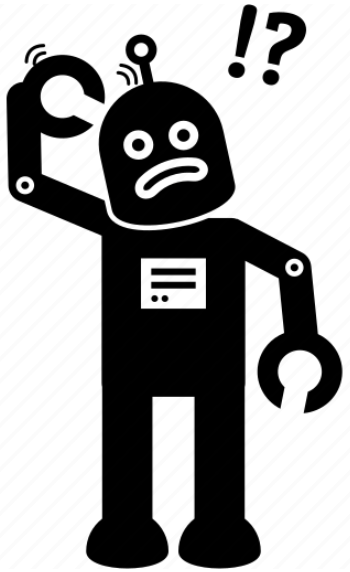
- **Issue**: The resulting belief state is the same for any t. We only know if we observe that we caught the plane afterwards. This not very helpful!

We need a way to specify how likely it is that we will end up at the airport after the action!

Example: Catching a Flight with Probabilities

Probabilities: Suppose the agent believes the following:

$$P(on\ time \mid A_{25}) = 0.04$$
$$P(on\ time \mid A_{90}) = 0.80$$
$$P(on\ time \mid A_{120}) = 0.99$$
$$P(on\ time \mid A_{1440}) = 0.9999$$



A probabilistic belief states as a probability distribution over states:

$$\text{Results}(\text{at home}, A_{90}) = \{\text{on time: } 0.8,\ \text{missed flight: } 0.2,\ \text{at home: } 0\}$$

	A_{20}	A_{90}	A_{120}	A_{1444}
$P(on\ time)$	0.04	0.8	0.99	0.9999
$P(missed\ flight)$	0.96	0.2	0.01	1E-04
$P(at\ home)$	0	0	0	0

Belief states

Which action should the agent choose?

Making a Decision Under Uncertainty

Given outcome probabilities, which action should the agent choose?

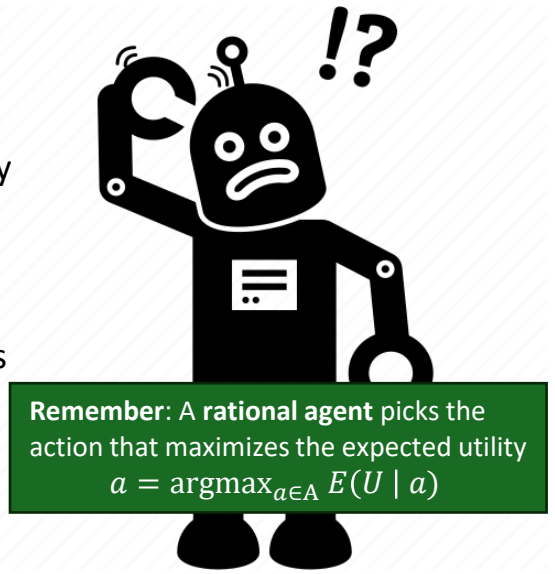
- Depends on **preferences** for missing a flight vs. time spent waiting.
- **Utility theory** represents preferences for different outcome using a utility function $U(outcome)$.

- **Decision Theory = Probability Theory + Utility Theory**
- The agent should choose actions that lead to an outcome that maximizes the **expected utility**.

$$\mathrm{a} = \mathrm{argmax}_{A_t}\, E[U(A_t)]$$

- The outcome depends on the action taken:

$$U(A_t) = U(reached\ state) - Cost(action)$$



Remember: A **rational agent** picks the action that maximizes the expected utility

$$a = \mathrm{argmax}_{a \in A}\, E(U \mid a)$$

Example:

Belief states (Probability Theory)

	A_{20}	A_{90}	A_{120}	A_{1444}
$P(on\ time)$	0.04	0.8	0.99	0.9999
$P(missed\ flight)$	0.96	0.2	0.01	1E-04
$P(at\ home)$	0	0	0	0

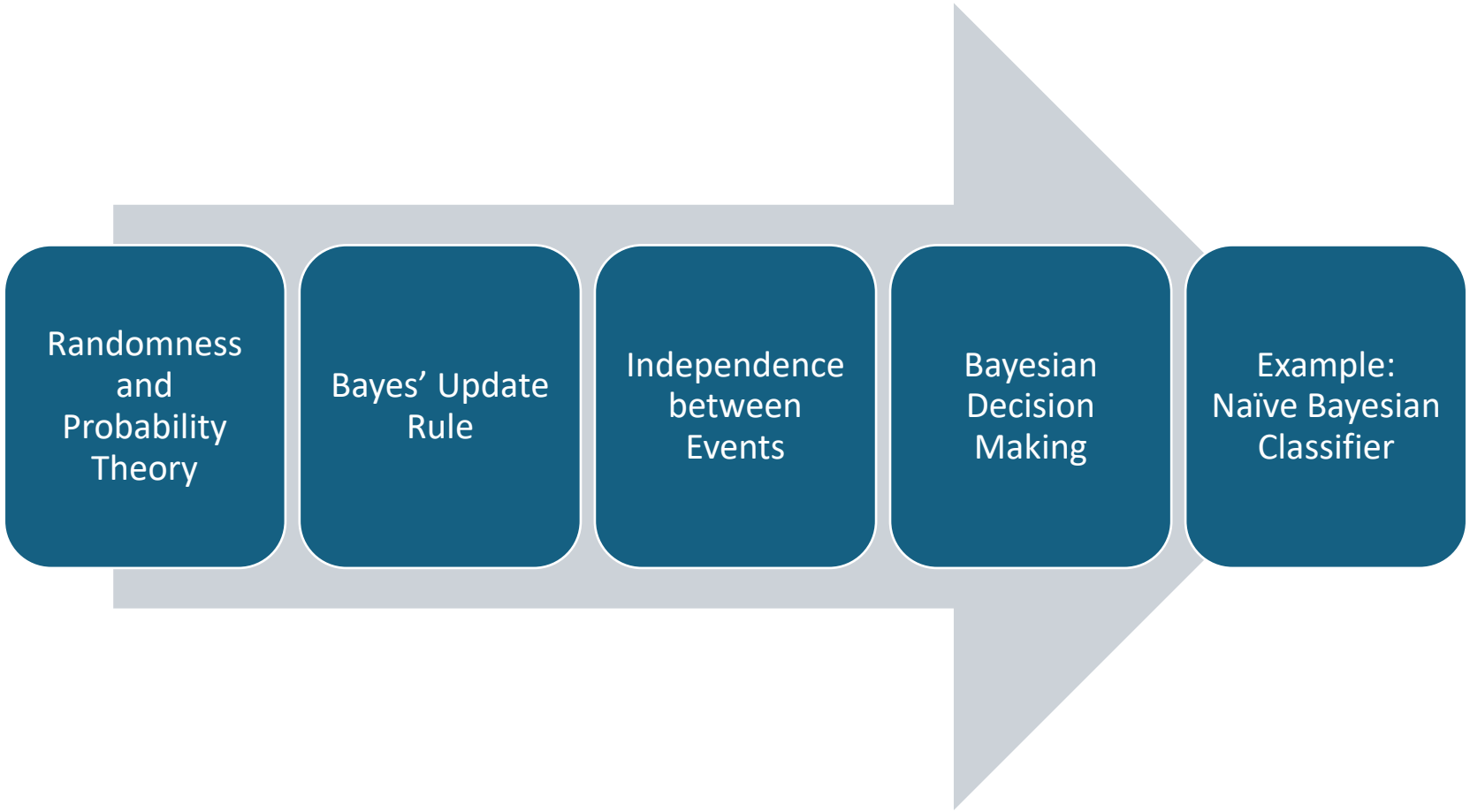
+

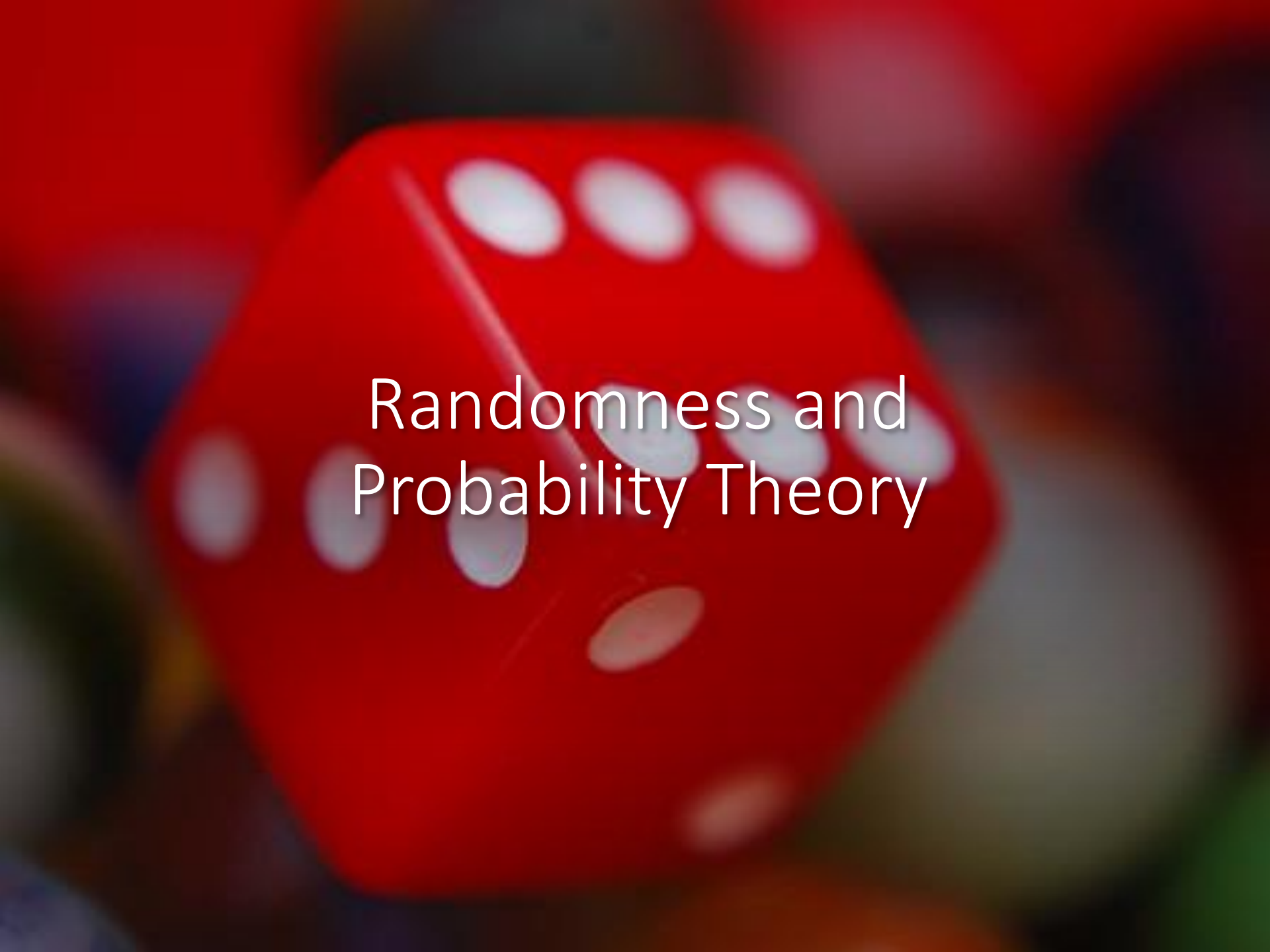
Utility structure

$U(on\ time)$	1000
$U(missed\ flight)$	-1000
$Cost(per\ minute)$	1

$$E[U(A_t)] = P(on\ time|A_t)\, U(on\ time \mid A_t) + P(missed\ flight|A_t) U(missed\ flight\ |A_t)$$
$$E[U(A_{120})] = 0.99 \times (1000 - 120 \times 1) + 0.01 \times (-1000 - 120 \times 1) = 860$$

Contents of this Module

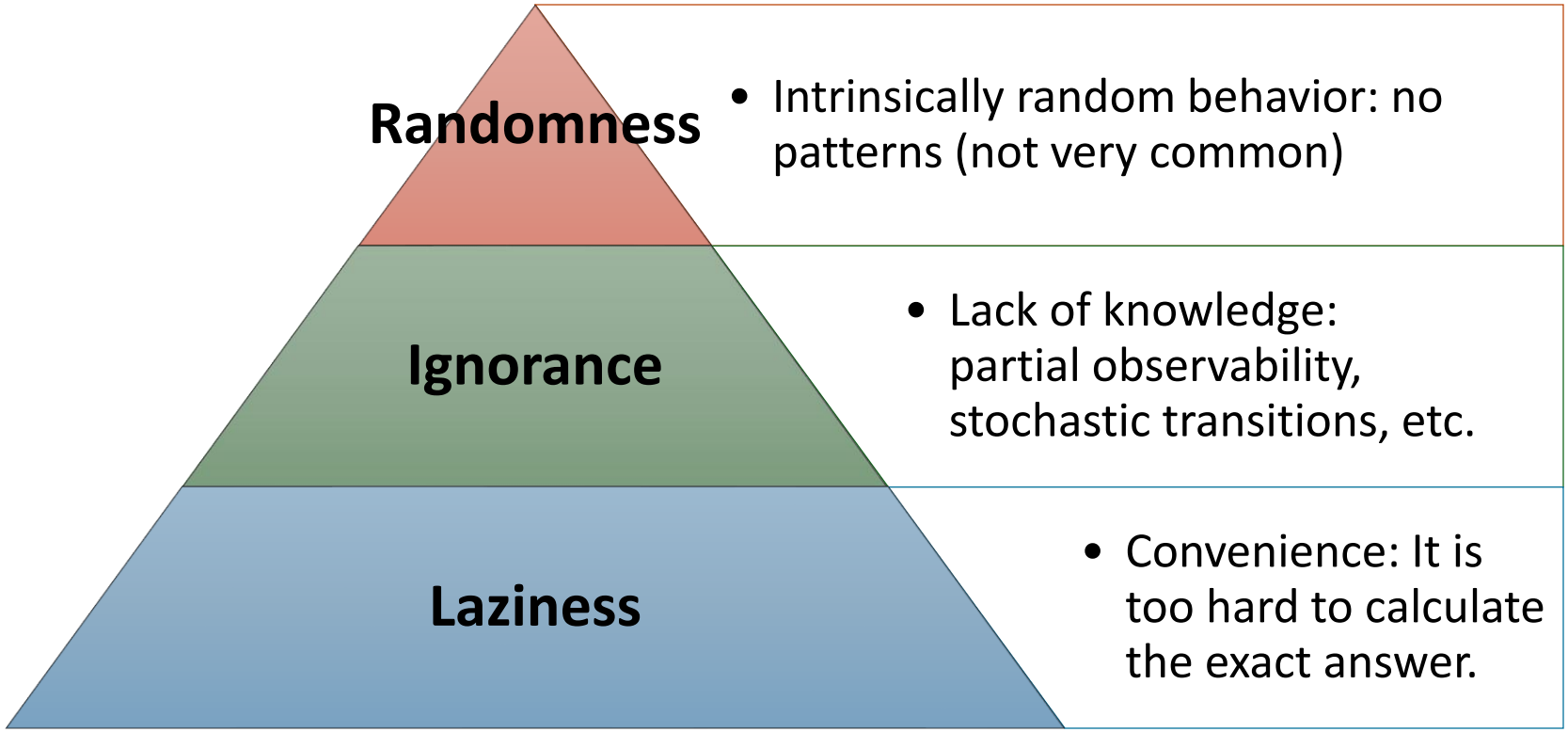




Randomness and Probability Theory

Sources of Uncertainty

Probabilistic assertions summarize effects of:



Example: What is the source of uncertainty for a coin toss?

What are Probabilities?

Frequentism (Objective; Positivist)

Probabilities are **long-run relative frequencies** determined by observation.

- For example, if we toss a coin **many times**, $P(heads)$ is estimated as the proportion of the time the coin will come up heads.
- But what if we are dealing with events that only happen once? E.g., what is the probability that a Republican will win the presidency in 2024? How do we define comparable elections? **Reference class problem**.

For lots of data

Bayesian Statistics (Subjective)

Probabilities are **degrees of belief** based on prior knowledge and updated by evidence.

Provides tools to:

- Assign belief values to statements without evidence
- Update our degrees of belief given observations = **Learning**

Limited data and learning

Both concepts are often used together.

Probability Theory Recap

- Notation: Prob. of an event $P(X = x) = P_X(x) = P(x)$
  Prob. distribution $\boldsymbol{P}(X) = \langle P(X = x_1), P(X = x_2), \dots, P(X = x_n) \rangle$

- Product rule $P(x, y) = P(x|y)P(y)$
- Chain rule $\boldsymbol{P}(\mathrm{X}_1, \mathrm{X}_2, \dots, \mathrm{X}_\mathrm{n}) = \mathbf{P}(\mathrm{X}_1)\mathbf{P}(\mathrm{X}_2|\mathrm{X}_1)\mathbf{P}(\mathrm{X}_3|\mathrm{X}_1, \mathrm{X}_2) \dots$
  $= \prod_{i=1}^{n} \boldsymbol{P}(X\_i|X_1, \dots, X_{i-1})$
- Conditional probability $P(x|y) = \frac{P(x,y)}{P(y)} = \alpha P(x, y)$
- Marginal distribution given $\boldsymbol{P}(X, Y)$ (joint probability distribution)
  $\boldsymbol{P}(X) = \sum_y \boldsymbol{P}(X, y)$ (marginalizing out Y)
- Independence
  - $X \perp\!\!\!\perp Y$: X, Y are independent (written as $X \perp\!\!\!\perp Y$) if and only if:
    $\forall x, y: P(x, y) = P(x)P(y)$
  - $X \perp\!\!\!\perp Y|Z$: X and Y are conditionally independent given Z if and only if:
    $\forall x, y, z: P(x, y|z) = P(x|z)P(y|z)$

Rev. Thomas Bayes
(1702-1761)

Bayesian Updates

Learning from Evidence

Bayes' Theorem: The Bayesian Update Rule

The product rule gives us two ways to factor a joint distribution for events $X = x$ and $E = e$:

$$P(x, e) = P(x \mid e)P(e) = P(e \mid x)P(x)$$

Therefore, $P(x \mid e) = \dfrac{P(e|x)\ P(x)}{P(e)}$

Posterior Prob.

Prior Prob.

Add evidence

Why is this useful?

- We can update our beliefs about an event x based on new evidence e.
- Update rule $P(x) \leftarrow \dfrac{P(e|x)\ P(x)}{P(e)}$

Written as distributions $\boldsymbol{P}(X \mid E) = \dfrac{\boldsymbol{P}(E|X)\boldsymbol{P}(X)}{\boldsymbol{P}(E)}$

Example: Getting Married in the Desert

New Evidence e

Prior probability of rain
$P(x)$ = 5/365 = 0.014

Marie is getting married tomorrow at an outdoor ceremony in the desert. In recent years, it has **rained only 5 days each year**. Unfortunately, the **weatherman has predicted rain** for tomorrow. When it actually rains, the weatherman **correctly forecasts rain 90%** of the time. When it doesn't rain, he incorrectly forecasts rain 10% of the time. What is Marie's belief for the **probability that it will rain** on her wedding day?

Likelihood
$P(e \mid x)$

Bayesian update: $P(x \mid e) = \frac{P(e|x)P(x)}{P(e)}$

Posterior Probability
$P(x \mid e)$?

$$P(\text{Rain}|\text{Predict}) = \frac{P(\text{Predict}|\text{Rain})P(\text{Rain})}{P(\text{Predict})}$$

$$= \frac{P(\text{Predict}|\text{Rain})P(\text{Rain})}{P(\text{Predict}|\text{Rain})P(\text{Rain}) + P(\text{Predict}|\neg\text{Rain})P(\neg\text{Rain})}$$

$$= \frac{0.9 \times 0.014}{0.9 \times 0.014 + 0.1 \times 0.986} = 0.111$$

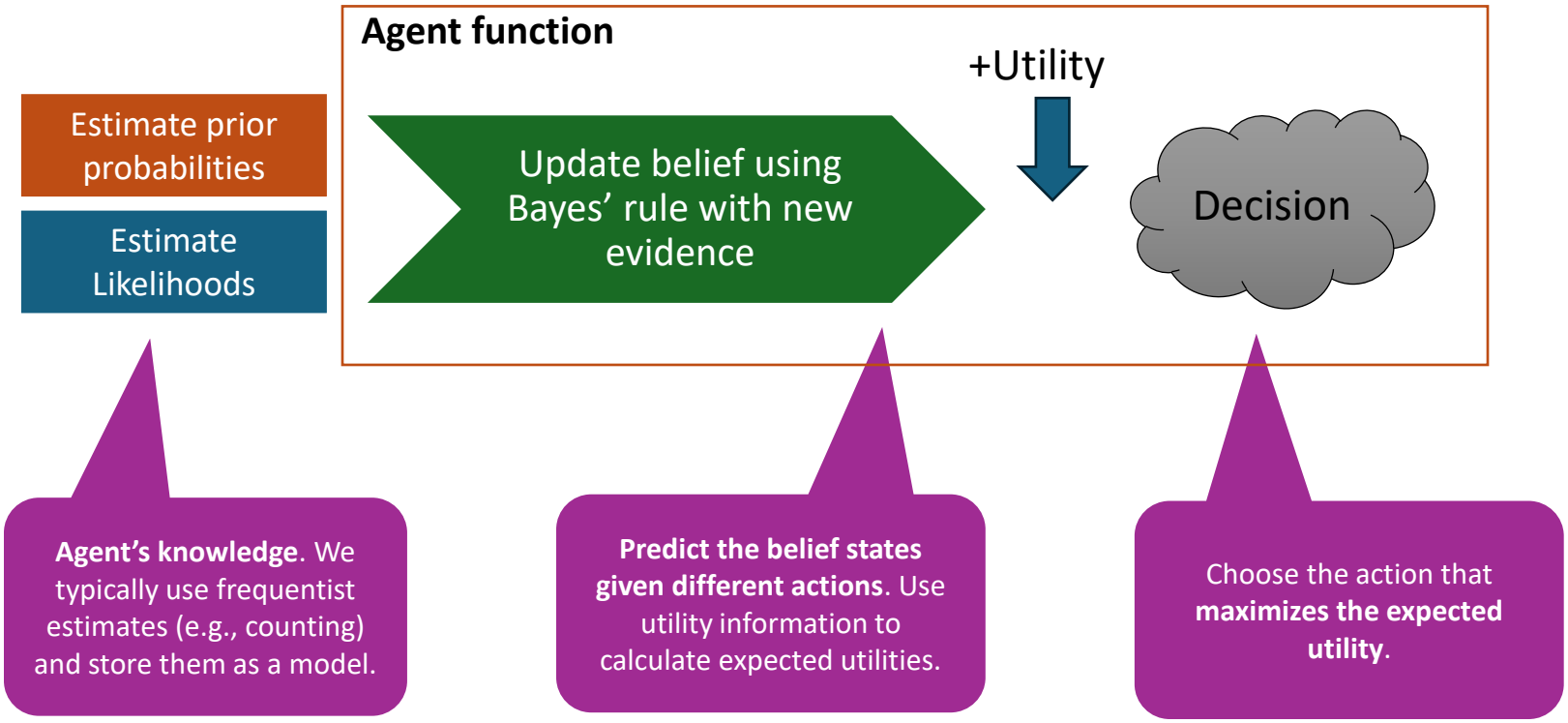
The weather forecast changes her belief from 0.014 to 0.111. She thinks that the chance of rain tomorrow is now about 10-times larger!

How much does Marie value no rain on her wedding day vs. moving the venue?

Bayesian Intelligent Agents

This is a type of utility-based agent, also called a decision-theoretic agent.

Approach




Independence
between Events

Issues with Making Decisions using Bayes' Rule

Approach

Issue: The table representing the likelihoods is typically way too large!

- For n random variables (evidence and outcome) with a domain size of k each, we have a table of size $O(k^n)$. This is a problem for
  - **storing** the table, and
  - **estimating** the probabilities from data (we need lots of data).

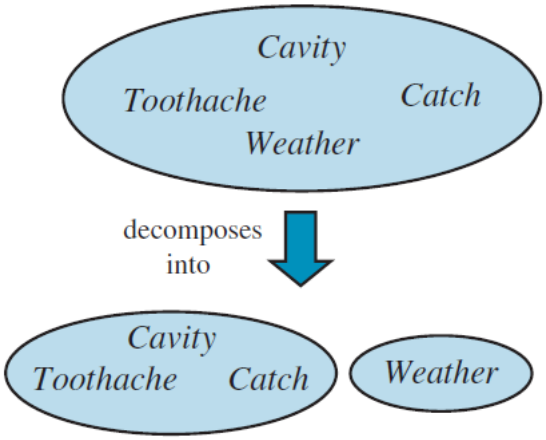
Solution:

- Decomposition of joint probability distributions using **conditional independence** between events.
- If we can identify conditional independence, then we can break the large table into several much smaller tables.

Independence Between Events

- Two events A and B are **independent** ($A \perp\!\!\!\perp B$) if and only if
$$\boldsymbol{P}(A, B) = \boldsymbol{P}(A)\,\boldsymbol{P}(B)$$

- This is equivalent to $\boldsymbol{P}(A \mid B) = \frac{\boldsymbol{P}(A,B)}{\boldsymbol{P}(B)} = \boldsymbol{P}(A)$ and $\boldsymbol{P}(B \mid A) = \boldsymbol{P}(B)$
- Independence is an important **simplifying assumption for modeling**. **Dentist Example**: *Cavity* and *Weather* can be assumed to be independent



Independence → **P**(Cavity, Weather) = **P**(Cavity)P(Weather)
P(Cavity | Weather) = **P**(Cavity)

Decomposition of the Joint Probability Distribution With Independence

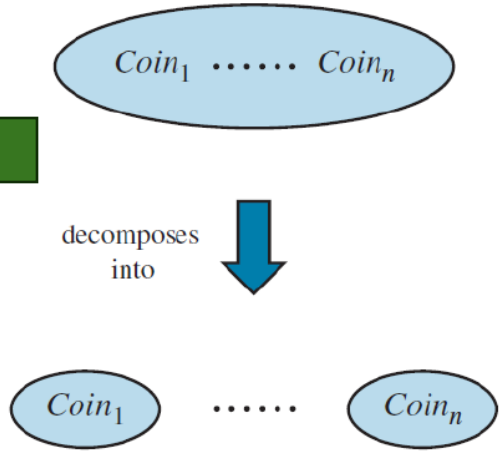
- **Independence:** The joint probability can be decomposed into

$2^n - 1$ entries

n entries

$$\boldsymbol{P}(Coin_1, \ldots, Coin_n) =$$

$$\boldsymbol{P}(Coin_1) \times \cdots \times \boldsymbol{P}(Coin_n) = \prod_{i=1}^{n} \boldsymbol{P}(Coin_i)$$



- The joint probability is a table with $\boldsymbol{2^n - 1}$ entries (all combinations of heads and tails).
- Independence reduces the numbers needed to specify the joint distribution to $\boldsymbol{n}$ probabilities (one for each coin).
- **Side note**: If we have identical (iid) coins, then we even only need 2 numbers, the probability of H and the number of coins.

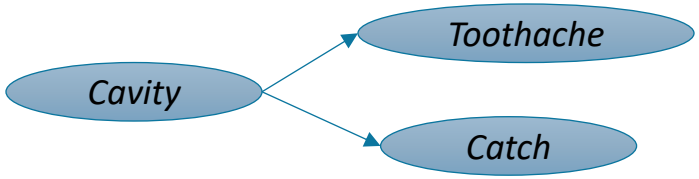
$O(2^n) \rightarrow O(n)$

Conditional Independence

- **Conditional independence**: A and B are *conditionally independent* given C (i.e., we know the value of C) if, and only if,

$$\boldsymbol{P}(A, B \mid C) = \boldsymbol{P}(A \mid C)\, \boldsymbol{P}(B \mid C)$$

Dentist Example:



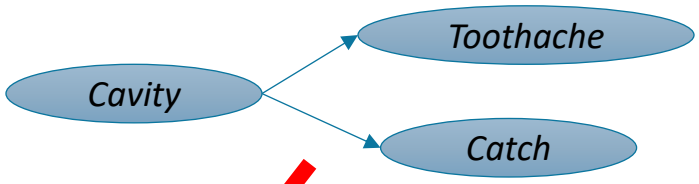
- The probability that the probe catches does not depend on whether he/she has a toothache if we know that the patient has a cavity:

  P(*Catch | Toothache, Cavity*) = **P**(*Catch | Cavity*)

- Therefore, *Catch* is conditionally independent of *Toothache* given *Cavity*

Decomposition of the Joint Probability Distribution With Conditional Independence

- **Conditional independence** simplifies the chain rule:



$2^3 - 1 =$ 7 entries

P(Toothache, Catch, Cavity) =
P(Cavity) **P**(Catch | Cavity) **P**(Toothache | ~~Catch~~, Cavity) =
P(Cavity) **P**(Catch | Cavity) **P**(Toothache | Cavity)

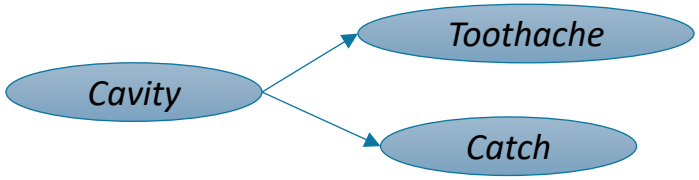
$1 + 2 + 2 =$ 5 entries

If each variable only depends on a small number of other variables:

$$O(2^n) \rightarrow O(n)$$

Bayesian Networks

Bayesian networks are a graphical method to specify dependence between random variables. This very useful technique will be discussed in detail later in this course.



- In many practical applications, each variable only depends on a small number of other variables.
- Conditional independence can **reduce the space requirements** to store the joint probability distribution from exponential to linear:

$$O(2^n) \rightarrow O(n)$$

- This means we can work efficiently with large models.

Bayesian Decision Making

Making Simple Decisions Under Uncertainty

Probabilistic Inference

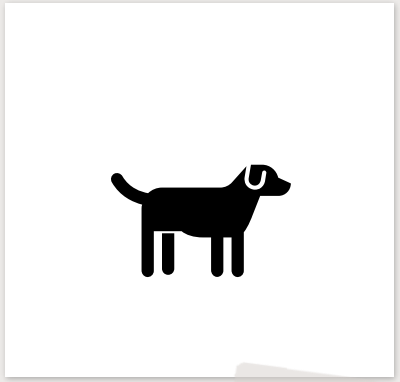
Suppose the agent must repeatedly guess the value of an unobserved *query variable* X given some observed *evidence* $E = e$ and we assume X probabilistically causes E.

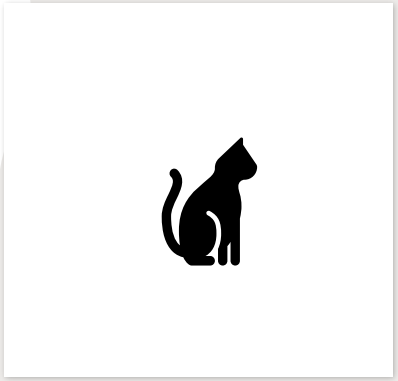
Example:

$\mathrm{x} \in \{\mathrm{dog}, \mathrm{zebra}, cat\}$, e = image features

What is the best guess $\hat{x}$?

Notation: We use $\hat{x}$ for an estimate and x^* for the optimal estimate.





The Optimal Decision Rule: MAP

- **Assumption:** The agent expresses the utility of the decision as a ***loss function***, which is 0 if the value of X is guessed correctly, and 1 otherwise.

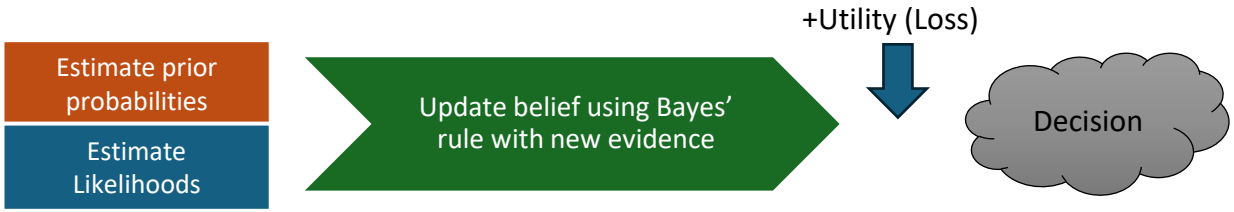
$$L(x, \hat{x}) = \begin{cases} 1 \text{ if } \hat{x} \neq x, \text{ and} \\ 0 \text{ otherwise.} \end{cases}$$

- The value for X that minimizes the ***expected loss*** is the one that has the greatest posterior probability given the evidence e.

$$\hat{x} = x^* = \text{argmax}_x \; P(X = x \mid E = e)$$

- This is called the **MAP** (maximum a posteriori) decision. Choosing the most likely x given e is **optimal for 0-1 loss**!
- The error of the Bayes decision rule is called the **Bayes Error Rate**. No classifier can do better!

MAP: Maximum A Posteriori Decision



0-1 loss means we should use the value x that has the highest (maximum) posterior probability given the evidence e, i.e., the prediction that most likely leads to a loss of 0.

Posterior Prob. Prior Prob.

$$x^* = \text{argmax}_x \, P(x|e) = \text{argmax}_x \frac{P(e|x)P(x)}{P(e)}$$

$$= \text{argmax}_x \, P(e|x)P(x)$$

$P(e)$ is fixed for a given evidence.

For comparison: the frequentist maximum likelihood decision ignores $P(x)$

$$x^* = \text{argmax}_x \, P(e|x)$$

Likelihood of observing e given class x

MAP: Example

We observe: e = stripes

What is the animal? $x \in \{\text{zebra, dog, cat}\}$

Posterior Prob.

$$x^* = \text{argmax}_x \, P(x|e) = \text{argmax}_x \frac{P(\text{stripes}|x)P(x)}{P(stripes)}$$

$$= \text{argmax}_x \, P(\text{stripes}|x)P(x)$$

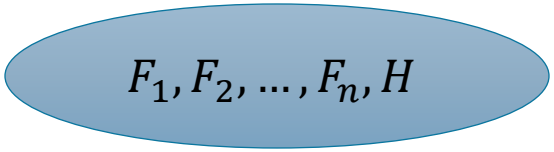
likelihood

Prior Prob.

Zebra: The likelihood $P(\text{stripes} \mid \text{zebra})$ is the highest. But the decision also depends on the prior $P(\text{zebra})$, the chance that we see a zebra.

Cat: The likelihood for cats having stripes may be smaller, but the prior probability of seeing a cat is much higher. Cat may have a larger posterior probability!

Bayes Classifier



- Suppose we have many different types of observations (evidence, symptoms, features) $F_1, \ldots, F_n$ that we want to use to decide on an underlying hypothesis H.

- The MAP decision involves estimating

$$h^* = \text{argmax}_{\text{h}\in H} \; P(f_1, \ldots, \; f_n|h)P(h)$$

- How many entries does the tables $P(f_1, \ldots, \; f_n|h)$ have?

Answer: If we assume that each feature can take on k values then the table has $\mathbf{O}(\boldsymbol{k^n})$ entries! What if we have 1000s of features?

Naïve Bayes Model

- We want to use the MAP decision which involves estimating

$$\mathrm{h}^* = \mathrm{argmax}_{\mathrm{h}\in H}\ P(f_1, \dots,\ f_n|h)P(h)$$

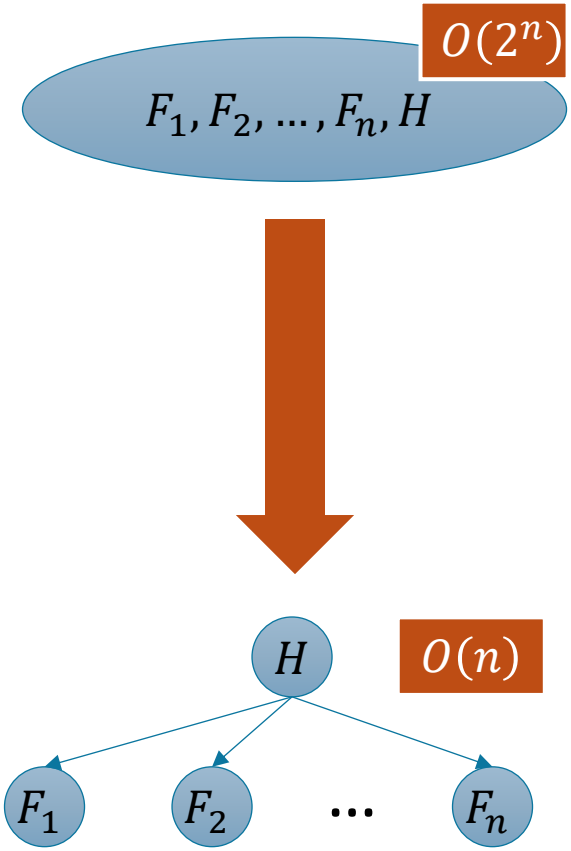
- **Issue**: The likelihood table size grows for n variables with k different values exponentially with $\mathrm{O}(k^n)$

- The naïve Bayes model makes the **simplifying assumption** that the different **features are conditionally independent *given the hypothesis.***

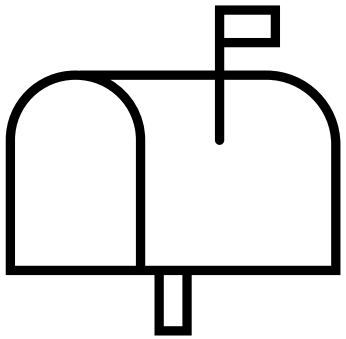
  This reduces the needed number of probabilities to $\mathrm{O}(k \times n)$:

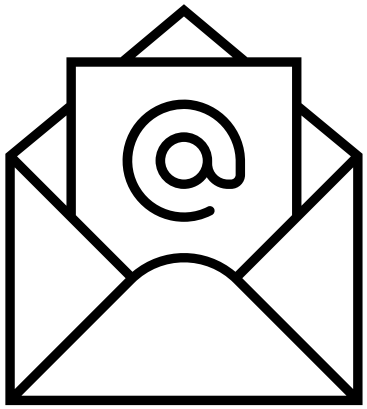
$$\hat{h} = \mathrm{argmax}_{\mathrm{h}\in H}\ P(h) \prod_{i=1}^{n} P(f_i|h)$$

- The naïve Bayes decision is not optimal.



Example: Naïve Bayesian Spam Filter





Example: Naïve Bayes Spam Filter

Dear Sir.

First, I must solicit your confidence in this transaction, this is by virture of its nature as being utterly confidencial and top secret. …

Ok, Iknow this is blatantly OT but I'm beginning to go insane. Had an old Dell Dimension XPS sitting in the corner and decided to put it to use, I know it was working pre being stuck in the corner, but when I plugged it in, hit the power nothing happened.

TO BE REMOVED FROM FUTURE MAILINGS, SIMPLY REPLY TO THIS MESSAGE AND PUT "REMOVE" IN THE SUBJECT.

99 MILLION EMAIL ADDRESSES FOR ONLY $99

Approach

To make decisions, we need to:

- Define random variables so we can estimate prior probabilities and likelihoods.
  - Class: spam no spam
  - Evidence: features of the message.
- Define utility/loss

Message Features: Bag of Words from Natural Language Processing (NLP)

Dear Sir.

First, I must solicit your confidence in this transaction, this is by virture of its nature as being utterly confidencial and top secret. …

TO BE REMOVED FROM FUTURE MAILINGS, SIMPLY REPLY TO THIS MESSAGE AND PUT "REMOVE" IN THE SUBJECT.

99 MILLION EMAIL ADDRESSES FOR ONLY $99

Ok, Iknow this is blatantly OT but I'm beginning to go insane. Had an old Dell Dimension XPS sitting in the corner and decided to put it to use, I know it was working pre being stuck in the corner, but when I plugged it in, hit the power nothing happened.

- Model a document as a vector of binary random variables $(W_1, \ldots, W_n)$.
- Each random variable represents if a specific word i is present $(W_i = 1)$ or not $(W_i = 0)$ in the message.
- Simplifications used by bag-of-words:
  - The order of the words in the message is ignored.
  - How often a word is repeated is ignored.
  - Uses a fixed vocabulary. Unknown words are ignored.

Naïve Bayes Spam Filter Using Words

- We model the **words** used in messages as **depending on the type of message** (h = spam or not spam), and we use the naïve simplifying assumption that **words are conditionally independent** given the type of message:

$$P(\text{message}|\text{h}) = P(w_1, \dots, w_n|\text{h}) \approx \prod_{i=1}^{n} P(w_i|\text{h})$$

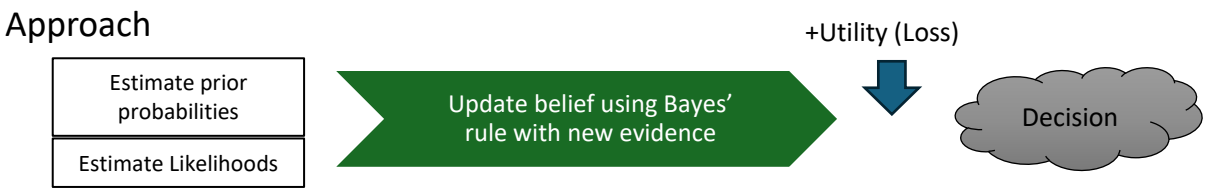
- Now we can calculate the a posteriori probability after the evidence of the message as

$$\underbrace{P(\text{h}|w_1, \dots, w_n)}_{\text{posterior}} \propto \underbrace{P(\text{h})}_{\text{prior}} \underbrace{\prod_{i=1}^{n} P(w_i|\text{h})}_{\text{likelihoods}}$$

likelihoods (presents and absence of words)

Note: It is only proportional since we do not divide by $P(w_1, \dots, w_n)$

Naïve Bayes Spam Filter: Decision Making



Update with words as evidence:

$$score(spam) = P(spam) \prod_{i=1}^{n} P(w_i|spam)$$

$$score(\neg spam) = P(\neg spam) \prod_{i=1}^{n} P(w_i|\neg spam)$$

A posteriori probabilities are simplified to proportional scores that can be compared

MAP Decision: $\hat{h} = \text{argmax}_h \ P(h|message)$

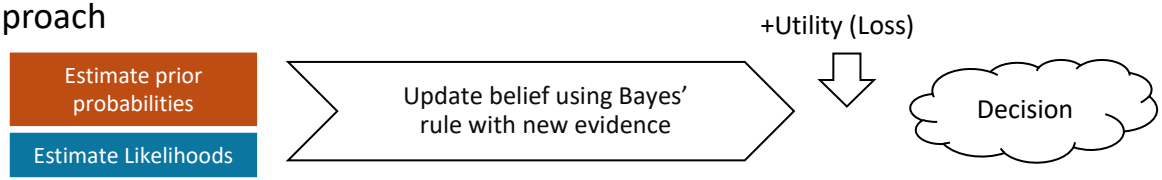
Minimizes 0-1 Loss (number of mistakes)

Scores are proportional to the probability. That means predict spam if

$$score(spam) > score(\neg spam)$$

Naïve Bayes Spam Filter: Parameter Estimation

Approach



Count in training data:

$$P(H = spam) = \frac{\#\ of\ spam\ messages + 1}{total\ \#\ of\ messages + \#\ of\ classes}$$

Smoothing for low counts.

$$P(w_i = 1 \mid H = spam) = \frac{\#\ of\ spam\ messages\ that\ contain\ the\ word + 1}{total\ \#\ of\ spam\ messages + \#\ of\ classes}$$

$O(n)$

Prior $\boldsymbol{P}(H)$

spam:	0.33
¬spam:	0.67

$P(W_i = 1 \mid H = spam)$

the :	0.0156
to :	0.0153
and :	0.0115
of :	0.0095
you :	0.0093
a :	0.0086
with:	0.0080
from:	0.0075
...	

$P(W_i = 1 \mid H = \neg spam)$

the :	0.0210
to :	0.0133
of :	0.0119
2002:	0.0110
with:	0.0108
from:	0.0107
and :	0.0105
a :	0.0100
...	

+ likelihoods for the absence of words:

$$P(W_i = 0 \mid H = spam) = 1 - P(W_i = 1 \mid H = spam)$$
$$P(W_i = 0 \mid H = \neg spam) = 1 - P(W_i = 1 \mid H = \neg spam)$$

Summary

Bayesian Intelligent Agent

- This is a type of utility-based agent, also known as a **decision-theoretic agent**.
- It combines:
  1. Probability theory to update its belief about outcomes given actions.
  2. Utility theory to represent preference for different outcomes.

- Bayes' Theorem provides a general framework for learning functions and decision rules from data is the goal of **Machine Learning**.
- An issue is that we need to estimate/learn a model consisting of an **exponentially large set of all likelihoods**. This is essentially the complete joint probability distribution between the evidence and state random variables!
- Much of AI and ML is about overcoming this model size issue by using simplifications, such as the naïve Bayes model.

Appendix: A Quick Review of Probability Theory

Random variables

Events

Joint probabilities

Marginal probabilities

Conditional probabilities

Random Variables

Random Variable

- We describe the (uncertain) state of the world using *random variables.*
- Random variables are denoted by capital letters.

- **R**: *Is it raining?*
- **W**: *What's the weather?*
- **Die**: *What is the outcome of rolling two dice?*
- **V**: *What is the speed of my car (in MPH)?*

Domain

- Random variables take on values in a *domain D.*
- Domain values must be mutually exclusive and exhaustive.

- **R** ∈ {True, False}
- **W** ∈ {Sunny, Cloudy, Rainy, Snow}
- **Die** ∈ {(1,1), (1,2), … (6,6)}
- **V** ∈ [0, 200]

Events and Propositions

Probabilistic statements are defined over **events**, world states or sets of states

- *"It is raining"*
- *"The weather is either cloudy or snowy"*
- *"The sum of the two dice rolls is 11"*
- *"My car is going between 30 and 50 miles per hour"*

Events are described using **propositions**:

- R = True
- W = "Cloudy" $\vee$ W = "Snowy"
- D $\in$ {(5,6), (6,5)}
- $30 \leq S \leq 50$

Probabilities

Probabilities are numbers indicating how likely we think an event (a realization of a random variable) is. These numbers can be

- Estimated as long-term averages (frequentist approach)
- Indicate a subjective belief (Bayesian approach)

Kolmogorov's 3 axioms are sufficient to define probability theory:

1. Probabilities are non-negative real numbers.
2. The probability that at least one atomic event happens is 1 (nothing happens is an event!).
3. The probability of mutually exclusive events is additive.

The axioms lead to important properties op probabilities (A and B are sets of events):

- Numeric bound: $0 \leq P(A) \leq 1$
- Monotonicity: if $A \subseteq B$ then $P(A) \leq P(B)$
- Addition law: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- Probability of the empty set: $P(\emptyset) = 0$
- Complement rule: $P(\neg A) = 1 - P(A)$

- Continuous variables need in addition the definition of density functions.

Joint Probability Distributions: Atomic Events

- **Atomic event**: a complete assignment of values to **all random variables**.
- **For AI**: Random variables are the fluents of a factored state description. An atomic event is a complete **specification of the state** of the world.

- Atomic events are mutually exclusive and exhaustive.

- **Example**: if the state consists of only two Boolean variables Cavity and Toothache, then there are 4 distinct atomic events:

  $Cavity = false \wedge Toothache = false$
  $Cavity = false \wedge Toothache = true$
  $Cavity = true \wedge Toothache = false$
  $Cavity = true \wedge Toothache = true$

Joint Probability Distributions

A **joint distribution** is an assignment of probabilities to every possible atomic event (state). The distribution is often stored as a table.

Example: Joint probability distribution for a world with two random variables

Atomic event	P
Cavity = false ∧Toothache = false	0.8
Cavity = false ∧ Toothache = true	0.1
Cavity = true ∧ Toothache = false	0.05
Cavity = true ∧ Toothache = true	0.05

Sum: 1.00

or

Cavity	Toothache	P
False	false	0.8
False	True	0.1
True	false	0.05
True	true	0.05

Notation:

- $P(X = x)$ or $P_X(x)$ or $P(x)$ for short, is the probability of the event that random variable X has taken on the value x.
- $\boldsymbol{P}(X)$ is the **distribution of probabilities** for all possible values of X. Often we are lazy or forget to make **P** bold.

Marginal Probability Distributions

Sometimes we are only interested in one variable (part of the state). This is called the *marginal distribution* $\boldsymbol{P}(Y)$

Joint Prob. Distr.

Cavity, Toothache	P
Cavity = false ∧Toothache = false	0.8
Cavity = false ∧ Toothache = true	0.1
Cavity = true ∧ Toothache = false	0.05
Cavity = true ∧ Toothache = true	0.05

Marginal Prob. Distr.

Cavity	P
Cavity = false	?
Cavity = true	?

Toothache	P
Toothache = false	?
Toothache = true	?

Marginal Probability Distributions 2

Suppose we have the joint distribution $\boldsymbol{P}(X, Y)$ and we want to find the *marginal distribution* $\boldsymbol{P}(X)$

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

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

Rule: to find $P(X = x)$, sum the probabilities of all atomic events where $X = x$.
This is called "**summing out**" or "**marginalizing out**" the other variables.

Marginal Probability Distributions 3

Suppose we have the joint distribution $\boldsymbol{P}(X, Y)$ and we want to find the *marginal distribution* $\boldsymbol{P}(Y)$.

Joint Prob. Distr.

Cavity, Toothache	P
Cavity = false ∧Toothache = false	0.8
Cavity = false ∧ Toothache = true	0.1
Cavity = true ∧ Toothache = false	0.05
Cavity = true ∧ Toothache = true	0.05

Marginal Prob. Distr.

Cavity	P
Cavity = false	0.8 + 0.1 = 0.9
Cavity = true	0.05 + 0.05=0.1

Toothache	P
Toothache = false	0.8 + 0.5= 0.85
Toothache = true	0.1 + 0.05= 0.15

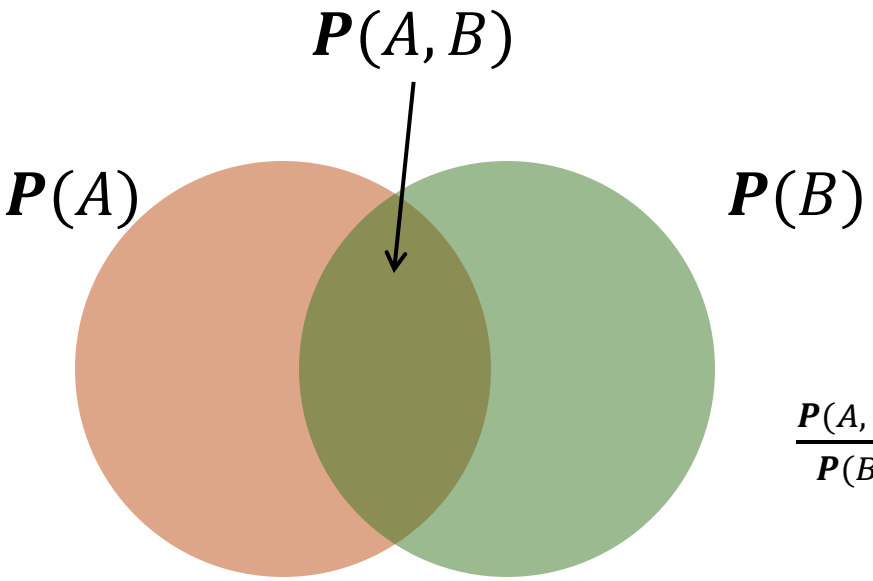
Conditional Probability

- Probability of event cavity given toothache:

$$\mathrm{P}(\mathrm{Cavity} = \mathrm{true} \mid \mathrm{Toothache} = \mathrm{true}) = \frac{\mathrm{P}(\mathrm{Cavity} = \mathrm{true} \wedge \mathrm{Toothache} = \mathrm{true})}{\mathrm{P}(\mathrm{Toothache} = \mathrm{true})}$$

- Conditional distribution of random variable A given B

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



$\frac{\boldsymbol{P}(A, B)}{\boldsymbol{P}(B)}$... fraction of B that is shared with A

Conditional Probability 2

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

Joint Prob. Distr.

Cavity, Toothache	P
Cavity = false ∧Toothache = false	0.8
Cavity = false ∧ Toothache = true	0.1
Cavity = true ∧ Toothache = false	0.05
Cavity = true ∧ Toothache = true	0.05

Marginal Prob. Distr.

Cavity	P
Cavity = false	0.9
Cavity = true	0.1

Toothache	P
Toothache = false	0.85
Toothache = true	0.15

What is P(*Cavity = true | Toothache = false*)?

$$\frac{P(\text{Cavity = true} \ \wedge \ \text{Toothache = false})}{P(\text{Toothache = false})} = 0.05 \ / \ 0.85 \ = \ 0.059$$

What is P(*Cavity = false | Toothache = true*)?

$$\frac{P(\text{Cavity = false} \ \wedge \ \text{Toothache = true})}{P(\text{Toothache = true})} = \ 0.1 \ / \ 0.15 \ = \ 0.667$$

Conditional Distributions

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

Joint Prob. Distr.

Cavity, Toothache	P
Cavity = false ∧Toothache = false	0.8
Cavity = false ∧ Toothache = true	0.1
Cavity = true ∧ Toothache = false	0.05
Cavity = true ∧ Toothache = true	0.05

A conditional distribution is a distribution over the values of one variable given fixed values of other variables.

Examples:

Conditional Prob. Distr.

Cavity | Toothache = true	P
Cavity = false | Toothache = true	0.667
Cavity = true | Toothache = true	0.333

Cavity | Toothache = false	P
Cavity = false | Toothache = false	0.941
Cavity = true | Toothache = false	0.059

Toothache | Cavity = true	P
Toothache= false | Cavity = true	0.5
Toothache = true | Cavity = true	0.5

Toothache | Cavity = false	P
Toothache= false | Cavity = false	0.889
Toothache = true | Cavity = false	0.111

Normalization Trick

To get the whole conditional distribution $\boldsymbol{P}(X \mid Y = y)$ at once, select all entries in the joint distribution matching $Y = y$ and renormalize them to sum to one.

Example: Calculate $\mathrm{P}(\mathrm{Toothache} \mid \mathrm{Cavity} = \mathrm{false})$ from the joint probability distribution:

Joint Prob. Distr.

Cavity, Toothache	P
Cavity = false ∧Toothache = false	**0.8**
Cavity = false ∧ Toothache = true	**0.1**
Cavity = true ∧ Toothache = false	0.05
Cavity = true ∧ Toothache = true	0.05

Select $P(X, Y = y)$

Toothache, Cavity = false	P
Toothache= false ∧ Cavity = false	0.8
Toothache = true ∧ Cavity = false	0.1

Sum is the marginal $P(Y = y) = 0.9$

Renormalize sum to 1 (= divide by $P(Y = y)$)

Cond. Prob. Distr.

Toothache | Cavity = false	P
Toothache= false | Cavity = false	0.889
Toothache = true | Cavity = false	0.111

Equivalent to

$$\boldsymbol{P}(X \mid Y = y) = \alpha\, \boldsymbol{P}(X, Y = y)$$

with $\alpha = 1/P(Y = y)$

Conditional Independence and the Bayes' Theorem

- These important concepts are introduced earlier in this module.

Conclusion

- Probability theory has many applications to deal with uncertainty in AI. Here are some examples:

  - **Joint probability distribution over the state space**: The basis of reasoning about how likely it is to be in a state.
  - **Conditional independence** makes it possible to work with more complicated factored state representations (a larger number of fluents).
  - **Marginal distributions** to reason about the most likely value of a fluent.
  - **Conditional distributions**: Transition probability given a chosen action.
  - **Bayesian updates**: Learn (= update the belief) about the current state.
  - **Machine learning** is based on decision theory, Bayesian decision making, and Bayesian updates for learning.