Foundations of Artificial Intelligence

11. Making Simple Decisions under Uncertainty
Probability Theory, Bayesian Networks, Other Approaches

Joschka Boedecker and Wolfram Burgard and
Frank Hutter and Bernhard Nebel and Michael Tangermann

Albert-Ludwigs-Universität Freiburg

June 19, 2019
Contents

1 Motivation

2 Foundations of Probability Theory

3 Probabilistic Inference

4 Bayesian Networks

5 Alternative Approaches
Lecture Overview

1. Motivation
2. Foundations of Probability Theory
3. Probabilistic Inference
4. Bayesian Networks
5. Alternative Approaches
Motivation

- In many cases, our knowledge of the world is **incomplete** (not enough information) or **uncertain** (sensors are unreliable).
- Often, rules about the domain are incomplete or even incorrect
  - e.g., *qualification problem*: what are the preconditions for an action?
- We have to act in spite of this!
- Drawing conclusions under **uncertainty**
Example

- **Goal**: Be in Freiburg at 9:15 to give a lecture.

- There are several **plans** that achieve the goal:
  - $P_1$: Get up at 7:00, take the bus at 8:15, the train at 8:30, arrive at 9:00 . . .
  - $P_2$: Get up at 6:00, take the bus at 7:15, the train at 7:30, arrive at 8:00 . . .
  - . . .

- All these plans are correct, but

  - They imply different **costs** and different **probabilities** of actually achieving the goal.

  - $P_2$ eventually is the plan of choice, since giving a lecture is very important, and the success rate of $P_1$ is only 90-95%.
Uncertainty in Logical Rules (1)

Example: Expert dental diagnosis system.

\[ \forall p [ \text{Symptom}(p, \text{toothache}) \Rightarrow \text{Disease}(p, \text{cavity}) ] \]

→ This rule is *incorrect*! Better:

\[ \forall p [ \text{Symptom}(p, \text{toothache}) \Rightarrow \text{Disease}(p, \text{cavity}) \lor \text{Disease}(p, \text{gum\textunderscore disease}) \lor \ldots ] \]

...however, we do not know all the causes.

Perhaps a *causal* rule is better?

\[ \forall p [ \text{Disease}(p, \text{cavity}) \Rightarrow \text{Symptom}(p, \text{toothache}) ] \]

→ Does not allow to reason from symptoms to causes & is still wrong!
We cannot enumerate all possible causes, and even if we could . . .

We do not know how correct the rules are (in medicine)

. . . and even if we did, there will always be uncertainty about the patient (the coincidence of having a toothache and a cavity that are unrelated, or the fact that not all tests have been run)

Without perfect knowledge, logical rules do not help much!
Let us suppose we wanted to support the localization of a robot with (constant) landmarks. With the availability of landmarks, we can narrow down on the area.

Problem: **Sensors** can be imprecise.

→ From the fact that a landmark was perceived, we cannot conclude with certainty that the robot is at that location.

→ The same is true when no landmark is perceived.

→ Only the probability increases or decreases.
Degree of Belief and Probability Theory

- We (and other agents) are convinced by facts and rules only up to a certain degree.

- One possibility for expressing the degree of belief is to use probabilities.

- Probabilities as frequencies / subjective beliefs
  - e.g., the agent is 90% (or 0.9) convinced by its sensor information means that it believes that in 9 out of 10 cases, the information is correct

- Probabilities quantify the uncertainty that stems from lack of knowledge.

- Probabilities are not to be confused with vagueness. The predicate tall is vague; the statement, “A man is 1.75–1.80m tall” is uncertain.
Uncertainty and Rational Decisions

- We have a choice of actions (or plans).
- These can lead to different results (worlds) with different probabilities.
- The actions have different (subjective) costs.
- The results have different (subjective) utilities.
- It would be rational to choose the action with the maximum expected total utility!

Decision Theory = Utility Theory + Probability Theory
function DT-AGENT(percept) returns an action

persistent: belief_state, probabilistic beliefs about the current state of the world
action, the agent’s action

update belief_state based on action and percept

calculate outcome probabilities for actions,

given action descriptions and current belief_state

select action with highest expected utility

given probabilities of outcomes and utility information

return action

**Decision theory:** An agent is rational exactly when it chooses the action with the maximum expected utility taken over all results of actions.
Lecture Overview

1 Motivation

2 Foundations of Probability Theory

3 Probabilistic Inference

4 Bayesian Networks

5 Alternative Approaches
A function $P$ mapping from formulae in propositional logic to the set $[0, 1]$ is a probability measure if for all propositions $\phi, \psi$ (whereby propositions are the equivalence classes formed by logically equivalent formulae):

1. $0 \leq P(\phi) \leq 1$
2. $P(\text{true}) = 1$
3. $P(\text{false}) = 0$
4. $P(\phi \lor \psi) = P(\phi) + P(\psi) - P(\phi \land \psi)$

All other properties can be derived from these axioms, for example:

$$P(\neg \phi) = 1 - P(\phi)$$

since $1 \overset{(2)}{=} P(\phi \lor \neg \phi) \overset{(4)}{=} P(\phi) + P(\neg \phi) - P(\phi \land \neg \phi) \overset{(3)}{=} P(\phi) + P(\neg \phi)$. 

(University of Freiburg)
Why are the Axioms Reasonable?

- If $P$ represents an *objectively* observable probability, the axioms clearly make sense.
- But why should an agent respect these axioms when it models its own degree of belief?

→ *Objective* vs. *subjective* probabilities

The axioms limit the set of beliefs that an agent can maintain.

One of the most convincing arguments for why subjective beliefs should respect the axioms was put forward by de Finetti in 1931. It is based on the connection between actions and degree of belief:

- If the beliefs do not follow the axioms, then there exists a betting strategy (the so-called “dutch book”) against the agent, where he will definitely loose!
We use random variables such as \( \text{Weather} \) (capitalized word), which has a domain of ordered values. In our case that could be \( \text{sunny}, \text{ rain}, \text{ cloudy}, \text{ snow} \) (lower case words).

A proposition might then be: \( \text{Weather} = \text{cloudy} \).

If the random variable is Boolean, e.g., \( \text{Headache} \), we may write either \( \text{Headache} = \text{true} \) or equivalently \( \text{headache} \) (lowercase!). Similarly, we may write \( \text{Headache} = \text{false} \) or equivalently \( \neg \text{headache} \).

Further, we can of course use Boolean connectors, e.g., \( \neg \text{headache} \wedge \text{Weather} = \text{cloudy} \).
Unconditional Probabilities (1)

\( P(a) \) denotes the \textit{unconditional} probability that it will turn out that \( A = \text{true} \) \textit{in the absence of any other information}, for example:

\[ P(\text{cavity}) = 0.1 \]
\[ P(\neg \text{cavity}) = 0.9 \]

In case of non-Boolean random variables:

\[ P(\text{Weather} = \text{sunny}) = 0.7 \]
\[ P(\text{Weather} = \text{rain}) = 0.2 \]
\[ P(\text{Weather} = \text{cloudy}) = 0.08 \]
\[ P(\text{Weather} = \text{snow}) = 0.02 \]
Unconditional Probabilities (2)

\( P(X) \) is the vector of probabilities for the (ordered) domain of the random variable \( X \):

\[
P(\text{Headache}) = \langle 0.1, 0.9 \rangle
\]

\[
P(\text{Weather}) = \langle 0.7, 0.2, 0.08, 0.02 \rangle
\]

define the probability distributions for the random variables \( \text{Headache} \) and \( \text{Weather} \).

\( P(\text{Headache} \land \text{Weather}) \) is a \( 4 \times 2 \) table of probabilities of all combinations of the values of a set of random variables.

<table>
<thead>
<tr>
<th>Headache = true</th>
<th>Headache = false</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{Weather} = \text{sunny} )</td>
<td>( P(\text{sunny} \land \text{headache}) )</td>
</tr>
<tr>
<td>( \text{Weather} = \text{rain} )</td>
<td></td>
</tr>
<tr>
<td>( \text{Weather} = \text{cloudy} )</td>
<td></td>
</tr>
<tr>
<td>( \text{Weather} = \text{snow} )</td>
<td></td>
</tr>
</tbody>
</table>
New information can change the probability.

Example: The probability of a cavity increases if we know the patient has a toothache.

If additional information is available, we can no longer use the prior probabilities!

\[ P(a \mid b) \] is the conditional or posterior probability of \( a \) given that all we know is \( b \):

\[ P(\text{cavity} \mid \text{toothache}) = 0.8 \]

\( P(X \mid Y) \) is the table of all conditional probabilities over all values of \( X \) and \( Y \).
**Conditional Probabilities (2)**

\( P(\text{Weather} \mid \text{Headache}) \) is a \( 4 \times 2 \) table of conditional probabilities of all combinations of the values of a set of random variables.

<table>
<thead>
<tr>
<th></th>
<th>Headache = true</th>
<th>Headache = false</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weather = sunny</td>
<td>( P(W = \text{sunny} \mid \text{headache}) )</td>
<td>( P(W = \text{sunny} \mid \neg \text{headache}) )</td>
</tr>
<tr>
<td>Weather = rain</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weather = cloudy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weather = snow</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Conditional probabilities result from unconditional probabilities (if \( P(b) > 0 \)) (by definition):

\[
P(a \mid b) = \frac{P(a \land b)}{P(b)}
\]

\[
P(a \land b) = P(a \mid b) \times P(b)
\]
P(X, Y) = P(X | Y)P(Y) corresponds to a system of equations:

\[ P(W = \text{sunny} \land \text{headache}) = P(W = \text{sunny} | \text{headache})P(\text{headache}) \]
\[ P(W = \text{rain} \land \text{headache}) = P(W = \text{rain} | \text{headache})P(\text{headache}) \]
\[ \ldots \]
\[ P(W = \text{snow} \land \neg \text{headache}) = P(W = \text{snow} | \neg \text{headache})P(\neg \text{headache}) \]
Product rule: \( P(a \land b) = P(a \mid b)P(b) \)

Similarly: \( P(a \land b) = P(b \mid a)P(a) \)

\( a \) and \( b \) are independent iff \( P(a \mid b) = P(a) \) (equiv. \( P(b \mid a) = P(b) \)).

Then (and only then) it holds that \( P(a \land b) = P(a)P(b) \).

Making this assumption: what is the effect wrt. computational efficiency?
Lecture Overview

1. Motivation

2. Foundations of Probability Theory

3. Probabilistic Inference

4. Bayesian Networks

5. Alternative Approaches
The agent assigns probabilities to every proposition in the domain. An atomic event assigns a value to every random variable \( X_1, \ldots, X_n \) (= complete specification of a state).

Example: Let \( X \) and \( Y \) be Boolean variables. Then we have the following 4 atomic events: \( x \land y, x \land \neg y, \neg x \land y, \neg x \land \neg y \).

The joint probability distribution \( P(X_1, \ldots, X_n) \) assigns a probability to every atomic event. Example of such a complete instantiation:

<table>
<thead>
<tr>
<th>cavity</th>
<th>( \text{toothache} )</th>
<th>( \neg \text{toothache} )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.04</td>
<td>0.06</td>
</tr>
<tr>
<td>( \neg \text{cavity} )</td>
<td>0.01</td>
<td>0.89</td>
</tr>
</tbody>
</table>

Observe: The sum of all fields is 1 (disjunction of events). Since all atomic events are disjoint, the conjunction of any two atomic events is necessarily false.
Working with the Joint Probability

All relevant probabilities can be computed using the joint probability by expressing them as a disjunction of atomic events.

Examples:

\[
P(\text{cavity} \lor \text{toothache}) = P(\text{cavity} \land \text{toothache}) + P(\neg \text{cavity} \land \text{toothache}) + P(\text{cavity} \land \neg \text{toothache})
\]

We obtain marginal probabilities by adding across a row or column:

\[
P(\text{cavity}) = P(\text{cavity} \land \text{toothache}) + P(\text{cavity} \land \neg \text{toothache})
\]

We obtain conditional probabilities by using a marginal probability:

\[
P(\text{cavity} \mid \text{toothache}) = \frac{P(\text{cavity} \land \text{toothache})}{P(\text{toothache})} = \frac{0.04}{0.04 + 0.01} = 0.80
\]
For any sets of variables $Y$ and $Z$ we have

$$P(Y) = \sum_z P(Y, z) = \sum_z P(Y \mid z) P(z)$$

law of total probability
Problems with Joint Probabilities

We can easily obtain all probabilities from the joint probability.

The joint probability, however, involves $k^n$ values, if there are $n$ random variables with $k$ values.

→ Difficult to represent
→ Difficult to assess

Questions:
→ Is there a more compact way of representing joint probabilities?
→ Is there an efficient method to work with this representation?

Answer: Not in general, but it can work in many cases. Modern systems work directly with conditional probabilities and make assumptions on the independence of variables (conditional independence) to simplify calculations.
Representing Joint Probabilities

Using the product rule $P(a \wedge b) = P(a | b) P(b)$, joint probabilities can be expressed as products of conditional probabilities.

$P(x_1, \ldots, x_n) = P(x_n, \ldots, x_1)$

Can these transformations change the required storage size?
Representing Joint Probabilities

Using the product rule $P(a \land b) = P(a \mid b)P(b)$, joint probabilities can be expressed as products of conditional probabilities.

$$P(x_1, \ldots, x_n) = P(x_n, \ldots, x_1) = P(x_n \mid x_{n-1}, \ldots, x_1)P(x_{n-1}, \ldots, x_1)$$
Using the product rule $P(a \land b) = P(a \mid b) P(b)$, joint probabilities can be expressed as products of conditional probabilities.

$$P(x_1, \ldots, x_n) = P(x_n, \ldots, x_1) = P(x_n \mid x_{n-1} \ldots, x_1) P(x_{n-1} \ldots, x_1)$$

$$= P(x_n \mid x_{n-1} \ldots, x_1) P(x_{n-1} \mid x_{n-2} \ldots, x_1) P(x_{n-2} \ldots, x_1) \cdots P(x_2 \mid x_1) P(x_1)$$
Using the product rule $P(a \land b) = P(a \mid b) P(b)$, joint probabilities can be expressed as products of conditional probabilities.

$$P(x_1, \ldots, x_n) = P(x_n, \ldots, x_1) = P(x_n \mid x_{n-1} \ldots, x_1) P(x_{n-1}, \ldots, x_1)$$

$$= P(x_n \mid x_{n-1} \ldots, x_1) P(x_{n-1} \mid x_{n-2} \ldots, x_1) P(x_{n-2}, \ldots, x_1)$$

$$= P(x_n \mid x_{n-1} \ldots, x_1) P(x_{n-1} \mid x_{n-2} \ldots, x_1) P(x_{n-2} \mid x_{n-3} \ldots, x_1)$$

$$P(x_{n-3}, \ldots, X_1)$$

$$= \ldots$$

$$= P(x_n \mid x_{n-1} \ldots, x_1) P(x_{n-1} \mid x_{n-2} \ldots, x_1) \ldots P(x_2 \mid x_1) P(x_1)$$

$$= \prod_{i=1}^{n} P(x_i \mid x_{i-1} \ldots x_1)$$

Can these transformations change the required storage size?
Bayes’ Rule

We know (product rule):

\[
P(a \land b) = P(a \mid b)P(b) \quad \text{and} \quad P(a \land b) = P(b \mid a)P(a)
\]

By equating the right-hand sides, we get

\[
P(a \mid b)P(b) = P(b \mid a)P(a)
\]

\[
\Rightarrow P(a \mid b) = \frac{P(b \mid a)P(a)}{P(b)}
\]

For multi-valued variables we get a set of equalities:

\[
P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)}
\]

Generalization (conditioning on background evidence \(e\)):

\[
P(Y \mid X, e) = \frac{P(X \mid Y, e)P(Y \mid e)}{P(X \mid e)}
\]
Applying Bayes’ Rule

\[ P(\text{toothache} \mid \text{cavity}) = 0.4 \]
\[ P(\text{cavity}) = 0.1 \]
\[ P(\text{toothache}) = 0.05 \]

\[ P(\text{cavity} \mid \text{toothache}) = \frac{0.4 \times 0.1}{0.05} = 0.8 \]

Why do we not try to assess \( P(\text{cavity} \mid \text{toothache}) \) directly?

\( P(\text{toothache} \mid \text{cavity}) \) (causal) is more robust than \( P(\text{cavity} \mid \text{toothache}) \) (diagnostic):

- \( P(\text{toothache} \mid \text{cavity}) \) is independent from the prior probabilities \( P(\text{toothache}) \) and \( P(\text{cavity}) \).
- If there is a cavity epidemic and \( P(\text{cavity}) \) increases, \( P(\text{toothache} \mid \text{cavity}) \) does not change, but \( P(\text{toothache}) \) and \( P(\text{cavity} \mid \text{toothache}) \) will change proportionally.
Relative Probability

Let’s say we would also like to consider the probability that our patient has gum disease.

\[
P(\text{toothache} | \text{gumdisease}) = 0.7
\]

\[
P(\text{gumdisease}) = 0.02
\]

Which diagnosis is more probable? Cavity or gum disease?

\[
P(c | t) = \frac{P(t | c)P(c)}{P(t)} \quad \text{or} \quad P(g | t) = \frac{P(t | g)P(g)}{P(t)}
\]

If we are only interested in the relative probability, we need not assess \(P(t)\):

\[
\frac{P(c | t)}{P(g | t)} = \frac{P(t | c)P(c)}{P(t)} \times \frac{P(t)}{P(t | g)P(g)} = \frac{P(t | c)P(c)}{P(t | g)P(g)}
\]

\[
= \frac{0.4 \times 0.1}{0.7 \times 0.02} = 2.857
\]

→ We elegantly excluded other possible diagnoses for toothache.
If we wish to determine the absolute probability of $P(c \mid t)$ but do not know $P(t)$, we can alternatively carry out a complete case analysis (e.g., for $c$ and $\neg c$) and use the fact that $P(c \mid t) + P(\neg c \mid t) = 1$ (here Boolean variables):
Normalization (2)

By substituting into the first equation:

\[
P(c \mid t) = \frac{P(t \mid c)P(c)}{P(t \mid c)P(c) + P(t \mid \neg c)P(\neg c)}
\]

For random variables with multiple values:

\[
P(Y \mid X) = \alpha P(X \mid Y)P(Y)
\]

where \(\alpha\) is the normalization constant needed to make the entries in \(P(Y \mid X)\) sum to 1 for each value of \(X\).

Example: \(\alpha(.1, .1, .3) = (.2, .2, .6)\).

Remark: In ML, relative probabilities often are sufficient.
Example

Your doctor tells you that you have tested positive for a serious but rare (1/10000) disease. This test ($t$) is correct to 99% (1% false positive & 1% false negative results).

What does this mean for you?

$$P(d | t) = \frac{P(t | d) \cdot P(d)}{P(t)}$$

Moral: If the test imprecision is much greater than the rate of occurrence of the disease, then a positive result is not as threatening as you might think.
Your doctor tells you that you have tested positive for a serious but rare (1/10000) disease. This test \((t)\) is correct to 99\% (1\% false positive & 1\% false negative results).

What does this mean for you?

\[
P(d \mid t) = \frac{P(t \mid d)P(d)}{P(t)} = \frac{P(t \mid d)P(d)}{P(t \mid d)P(d) + P(t \mid \neg d)P(\neg d)}
\]
Your doctor tells you that you have tested positive for a serious but rare (1/10000) disease. This test \((t)\) is correct to 99% (1% false positive & 1% false negative results).

What does this mean for you?

\[
P(d \mid t) = \frac{P(t \mid d)P(d)}{P(t)} = \frac{P(t \mid d)P(d)}{P(t \mid d)P(d) + P(t \mid \neg d)P(\neg d)}
\]

\[
P(d) = 0.0001 \quad P(t \mid d) = 0.99 \quad P(t \mid \neg d) = 0.01
\]

\[
P(d \mid t) = \frac{0.99 \times 0.0001}{0.99 \times 0.0001 + 0.01 \times 0.9999} = \frac{0.000099}{0.000099 + 0.009999} \\
= \frac{0.000099}{0.010088} \approx 0.01
\]

**Moral:** If the test imprecision is much greater than the rate of occurrence of the disease, then a positive result is not as threatening as you might think.
A probe by the dentist catches \( (\text{Catch} = \text{true}) \) in the aching tooth \( (\text{Toothache} = \text{true}) \) of a patient. We already know that \( P(\text{cavity} | \text{toothache}) = 0.8 \). Furthermore, using Bayes’ rule, we can calculate:

\[
P(\text{cavity} | \text{catch}) = 0.95
\]

But how does the combined evidence \( (\text{tooth} \land \text{catch}) \) help?

Using Bayes’ rule, the dentist could establish:

\[
P(\text{cav} | \text{tooth} \land \text{catch}) = \frac{P(\text{tooth} \land \text{catch} | \text{cav}) \times P(\text{cav})}{P(\text{tooth} \land \text{catch})} = \alpha P(\text{tooth} \land \text{catch} | \text{cav}) \times P(\text{cav})
\]
Problem: The dentist needs $P(\text{tooth} \land \text{catch} | \text{cav})$, i.e., diagnostic knowledge of all combinations of symptoms in the general case.

It would be nice if tooth and catch were independent but they are not: $P(\text{tooth} | \text{catch}) \neq P(\text{tooth})$ - if a probe catches in the tooth, it probably has a cavity which probably causes toothache.
Multiple Evidence (2)

Problem: The dentist needs $P(\text{tooth} \land \text{catch} \mid \text{cav})$, i.e., diagnostic knowledge of all combinations of symptoms in the general case.

It would be nice if \textit{tooth} and \textit{catch} were independent but they are not: $P(\text{tooth} \mid \text{catch}) \neq P(\text{tooth})$ - if a probe catches in the tooth, it probably has a cavity which probably causes toothache.

They are \underline{conditionally} independent given that we know whether the tooth has a cavity:

$$P(\text{tooth} \mid \text{catch}, \text{cav}) = P(\text{tooth} \mid \text{cav})$$

If one already knows that there is a cavity, then the additional knowledge of the probe catching does not change the probability.

$$P(\text{tooth} \land \text{catch} \mid \text{cav}) = P(\text{tooth} \mid \text{catch}, \text{cav})P(\text{catch} \mid \text{cav}) = P(\text{tooth} \mid \text{cav})P(\text{catch} \mid \text{cav})$$
Thus our diagnostic problem turns into:

\[ P(\text{cav} \mid \text{tooth} \land \text{catch}) = \alpha P(\text{tooth} \land \text{catch} \mid \text{cav}) P(\text{cav}) \]
Thus our diagnostic problem turns into:

\[ P(\text{cav} \mid \text{tooth} \land \text{catch}) = \alpha P(\text{tooth} \land \text{catch} \mid \text{cav})P(\text{cav}) \]

\[ = \alpha P(\text{tooth} \mid \text{catch}, \text{cav})P(\text{catch} \mid \text{cav})P(\text{cav}) \]
Thus our diagnostic problem turns into:

\[ P(\text{cav} \mid \text{tooth} \land \text{catch}) = \alpha P(\text{tooth} \land \text{catch} \mid \text{cav}) P(\text{cav}) \]

\[ = \alpha P(\text{tooth} \mid \text{catch, cav}) P(\text{catch} \mid \text{cav}) P(\text{cav}) \]

\[ = \alpha P(\text{tooth} \mid \text{cav}) P(\text{catch} \mid \text{cav}) P(\text{cav}) \]
Thus our diagnostic problem turns into:

\[ P(\text{cav} \mid \text{tooth} \land \text{catch}) = \alpha P(\text{tooth} \land \text{catch} \mid \text{cav}) P(\text{cav}) \]

\[ = \alpha P(\text{tooth} \mid \text{catch}, \text{cav}) P(\text{catch} \mid \text{cav}) P(\text{cav}) \]

\[ = \alpha P(\text{tooth} \mid \text{cav}) P(\text{catch} \mid \text{cav}) P(\text{cav}) \]

The general definition of conditional independence of two variables \(X\) and \(Y\) given a third variable \(Z\) (a common cause) is:

\[ P(X, Y \mid Z) = P(X \mid Z) P(Y \mid Z) \]
Eating icecream and observing sunshine is not independent

\[ P(\text{ice} \mid \text{sun}) \neq P(\text{ice}) \]

The variables *Ice* and *Sun* are not independent.
But if the reason for eating icecream is simply that it is hot outside, then the additional observation of sunshine does not make a difference:

\[ P(\text{ice} \mid \text{sun}, \text{hot}) = P(\text{ice} \mid \text{hot}) \]

The variables *Ice* and *Sun* are conditionally independent given that \( \text{Hot} = \text{true} \) is observed.
The knowledge about independence often comes from insight of the domain and is part of the modelling of the problem. Conditional independence can often be exploited to make things simpler (see later).
Problem: we would like to avoid calculating the full joint probability table!

Assuming conditional independence, multiple evidence can be reduced to prior probabilities and conditional probabilities. The general combination rule, if $Z_1$ and $Z_2$ are independent given $X$ is

$$P(X \mid Z_1, Z_2) = \alpha P(X)P(Z_1 \mid X)P(Z_2 \mid X)$$

where $\alpha$ is the normalization constant.
Recursive Bayesian Updating

Problem: we would like to avoid calculating the full joint probability table!

Assuming conditional independence, multiple evidence can be reduced to prior probabilities and conditional probabilities.

The general combination rule, if $Z_1$ and $Z_2$ are independent given $X$ is

$$P(X \mid Z_1, Z_2) = \alpha P(X) P(Z_1 \mid X) P(Z_2 \mid X)$$

where $\alpha$ is the normalization constant.

Generalization: Recursive Bayesian Updating

$$P(X \mid Z_1, \ldots, Z_n) = \alpha P(X) \prod_{i=1}^{n} P(Z_i \mid X)$$
Types of Variables

- Variables can be discrete or continuous:

  - Discrete variables
    - Weather: sunny, rain, cloudy, snow
    - Cavity: true, false (Boolean)

  - Continuous variables
    - Tomorrow’s maximum temperature in Freiburg
    - Domain can be the entire real line or any subset.
    - Distributions for continuous variables are typically given by probability density functions.
Uncertainty is unavoidable in complex, dynamic worlds in which agents are ignorant.

Probabilities express the agent’s inability to reach a definite decision. They summarize the agent’s beliefs.

Conditional and unconditional probabilities can be formulated over propositions.

If an agent disrespects the theoretical probability axioms, it is likely to demonstrate irrational behaviour.

Bayes’ rule allows us to calculate known probabilities from unknown probabilities.

Multiple evidence (assuming independence) can be effectively incorporated using recursive Bayesian updating.
Lecture Overview

1. Motivation

2. Foundations of Probability Theory

3. Probabilistic Inference

4. Bayesian Networks

5. Alternative Approaches
Example domain: I am at work. My neighbour John calls me to tell me, that my alarm is ringing. My neighbour Mary doesn’t call. Sometimes, the alarm is started by a slight earthquake.

Question: Is there a burglary?

Variables: Burglary, Earthquake, Alarm, JohnCalls, MaryCalls.
Domain knowledge/ assumptions:

- Events *Burglary* and *Earthquake* are independent. (of course, to be discussed: a burglary does not cause an earthquake, but a burglar might use an earthquake to do the burglary. Then the independence assumption is not true. This is a design decision!)

- *Alarm* might be activated by burglary or earthquake

- John calls if and only if he heard the alarm. His call probability is not influenced by the fact, that there is an earthquake at the same time. Same for Mary.

How to model this domain efficiently? Goal: Answer questions.
Bayesian Networks

(also belief networks, probabilistic networks, causal networks)

- The *random variables* are the *nodes*.
- **Directed edges** between nodes represent *direct influence*.
- A *table of conditional probabilities* (CPT) is associated with every node, in which the effect of the *parent* nodes is quantified.
- The graph is **acyclic** (a DAG).

Remark: *Burglary and Earthquake* are denoted as the *parents* of *Alarm*
The Meaning of Bayesian Networks

- **Alarm** depends on **Burglary** and **Earthquake**.
- **MaryCalls** only depends on **Alarm**.

\[
P(\text{maryCalls} \mid \text{alarm, burglary}) = P(\text{maryCalls} \mid \text{alarm}) \text{ and } P(\text{maryCalls} \mid \text{alarm, burglary, johnCalls, earthquake}) = P(\text{maryCalls} \mid \text{alarm})
\]

→ Bayesian Networks can be considered as sets of (conditional) independence assumptions.
Bayesian networks can be seen as a more compact representation of joint probabilities.

Let all nodes $X_1, \ldots, X_n$ be ordered topologically according to the arrows in the network. Let $x_1, \ldots, x_n$ be the values of the variables. Then

$$P(x_1, \ldots, x_n) = P(x_n \mid x_{n-1}, \ldots, x_1) \cdot \ldots \cdot P(x_2 \mid x_1)P(x_1)$$

$$= \prod_{i=1}^{n} P(x_i \mid x_{i-1}, \ldots, x_1)$$

According to the independence assumption, this is equivalent to

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i \mid \text{parents}(x_i))$$

We can calculate the joint probability from the network topology and the conditional probability tables (CPTs)!
Only prob. for pos. events are given, negative: \( P(\neg x) = 1 - P(x) \). Note: the size of the table depends on the number of parents!

\[
P(j, m, a, \neg b, \neg e) = \\
P(j | m, a, \neg b, \neg e)P(m | a, \neg b, \neg e)P(a | \neg b, \neg e)P(\neg b | \neg e)P(\neg e) \\
= P(j | a)P(m | a)P(a | \neg b, \neg e)P(\neg b)P(\neg e) \\
= 0.9 \times 0.7 \times 0.001 \times 0.999 \times 0.998 = 0.00062
\]
For the explicit representation of Bayesian networks, we need a table of size $2^n$ where $n$ is the number of variables.

In the case that every node in a network has at most $k$ parents, we only need $n$ tables of size $2^k$ (assuming Boolean variables).

**Example**: $n = 20$ and $k = 5$

$2^{20} = 1,048,576$ and $20 \times 2^5 = 640$ different explicitly-represented probabilities!

In the worst case, a Bayesian network can become exponentially large, for example if every variable is directly influenced by all the others.

The size depends on the application domain (local vs. global interaction) and the skill of the designer.
Naive Design of a Network

- Order all variables
- Take the first from those that remain
- Assign all direct influences from nodes already in the network to the new node (Edges + CPT).
- If there are still variables in the list, repeat from step 2.
Example 1

\[ M, J, A, B, E \]
Example 2

$M, J, E, B, A$
left = M, J, A, B, E, right = M, J, E, B, A

→ Appears to be an attempt to build a diagnostic model of symptoms and causes, which always leads to dependencies between causes that are actually independent and symptoms that appear separately.
Inference in Bayesian Networks

Instantiating evidence variables and sending queries to nodes.

What is \( P(\text{burglary} \mid \text{johnCalls}) \) or \( P(\text{burglary} \mid \text{johnCalls}, \text{maryCalls}) \)?
A node is conditionally independent of its non-descendants given its parents.
*JohnCalls* is independent of *Burglary* and *Earthquake* given the value of *Alarm*.
A node is conditionally independent of all other nodes in the network given the Markov blanket, i.e., its parents, children and children’s parents.
Burglary is independent of JohnCalls and MaryCalls, given the values of Alarm and Earthquake, i.e.,

\[
P(\text{Burglary} \mid \text{JohnCalls, MaryCalls, Alarm, Earthquake}) = P(\text{Burglary} \mid \text{Alarm, Earthquake})
\]
Exact Inference in Bayesian Networks

- Compute the **posterior probability distribution** for a set of query variables $X$ given an observation, i.e., the values of a set of evidence variables $E$.

- Complete set of variables is $X \cup E \cup Y$

- $Y$ are called the **hidden variables**

- Typical query $P(X \mid e)$ where $e$ are the observed values of $E$.

- In the remainder: $X$ is a singleton

Example:

$$P(Burglary \mid JohnCalls = true, MaryCalls = true) = (0.284, 0.716)$$
Inference by Enumeration

- \( P(X \mid e) = \alpha P(X, e) = \sum_y \alpha P(X, e, y) \)

- The network gives a complete representation of the full joint distribution.

- A query can be answered using a Bayesian network by computing sums of products of conditional probabilities from the network.

- We sum over the hidden variables.
Example

- Consider $P(Burglary \mid JohnCalls = true, MaryCalls = true)$
- The evidence variables are
Example

- Consider \( P(Burglary \mid JohnCalls = true, MaryCalls = true) \)
- The evidence variables are \( JohnCalls \) and \( MaryCalls \).
- The hidden variables are
Example

- Consider $P(Burglary \mid JohnCalls = true, MaryCalls = true)$
- The evidence variables are $JohnCalls$ and $MaryCalls$.
- The hidden variables are $Earthquake$ and $Alarm$.
- We have: $P(B \mid j, m) = \alpha P(B, j, m)$
Example

- Consider \( P(\text{Burglary} \mid \text{JohnCalls} = \text{true}, \text{MaryCalls} = \text{true}) \)

- The evidence variables are \( \text{JohnCalls} \) and \( \text{MaryCalls} \).

- The hidden variables are \( \text{Earthquake} \) and \( \text{Alarm} \).

- We have: \( P(B \mid j, m) = \alpha P(B, j, m) \)
  \[
  = \alpha \sum_e \sum_a P(B, j, m, e, a)
  \]

- If we consider the independence of variables, we obtain for \( B = \text{true} \)
  \[
P(b \mid j, m) = \alpha \sum_e \sum_a P(j \mid a)P(m \mid a)P(a \mid e, b)P(e)P(b)
  \]

- Reorganization of the terms yields:
  \[
P(b \mid j, m) = \alpha P(b)\sum_e P(e)\sum_a P(a \mid e, b)P(j \mid a)P(m \mid a)
  \]
Recall Bayesian Network for Domain

- **Burglary** with conditional probability $P(B) = 0.01$
- **Earthquake** with conditional probability $P(E) = 0.02$
- **Alarm**
  - Table: $\begin{array}{c|c|c}
  B & E & P(A) \\
  \hline
  T & T & 0.95 \\
  T & F & 0.94 \\
  F & T & 0.29 \\
  F & F & 0.001 \\
  \end{array}$
- **JohnCalls** with $P(J) = T: 0.90, F: 0.05$
- **MaryCalls** with $P(M) = T: 0.70, F: 0.01$
Evaluation of $P(b \mid j, m)$

$$P(b \mid j, m) = \alpha P(b) \sum_e P(e) \sum_a P(a \mid e, b) P(j \mid a) P(m \mid a)$$

$$\mathbf{P}(B \mid j, m) = \alpha (0.0006, 0.0015) = (0.284, 0.716)$$
function ENUMERATION-ASK($X$, $e$, $bn$) returns a distribution over $X$
inputs: $X$, the query variable
        $e$, observed values for variables $E$
        $bn$, a Bayes net with variables $\{X\} \cup E \cup Y$ /* $Y$ = hidden variables */

$Q(X) \leftarrow$ a distribution over $X$, initially empty
for each value $x_i$ of $X$ do
    $Q(x_i) \leftarrow$ ENUMERATE-ALL($bn$.VARS, $e_{x_i}$)
        where $e_{x_i}$ is $e$ extended with $X = x_i$
return NORMALIZE($Q(X)$)

function ENUMERATE-ALL($vars$, $e$) returns a real number
if EMPTY?($vars$) then return 1.0
$Y \leftarrow$ FIRST($vars$)
if $Y$ has value $y$ in $e$
    then return $P(y | parents(Y)) \times$ ENUMERATE-ALL(REST($vars$), $e$)
else return $\sum_y P(y | parents(Y)) \times$ ENUMERATE-ALL(REST($vars$), $e_y$)
    where $e_y$ is $e$ extended with $Y = y$
Properties of the **Enumeration-Ask Algorithm**

- The **Enumeration-Ask** algorithm evaluates the trees in a depth-first manner.

- Space complexity is linear in the number of variables.

- Time complexity for a network with $n$ Boolean variables is $O(2^n)$, since in the worst case, all terms must be evaluated for the two cases ("true" and "false")
Variable Elimination

- The enumeration algorithm can be improved significantly by eliminating repeating or unnecessary calculations.
- The key idea is to evaluate expressions from right to left (bottom-up) and to save results for later use.
- Additionally, unnecessary expressions can be removed.
Example

Let us consider the query $P(\text{JohnCalls} \mid \text{Burglary} = \text{true})$.

The nested sum is

$$P(j, b) = \alpha P(b) \sum_e P(e) \sum_a P(a \mid b, e) P(j, a) \sum_m P(m \mid a)$$

Obviously, the rightmost sum equals 1 so that it can safely be dropped.

A general observation: variables, that are not query or evidence variables and not ancestor nodes of query or evidence variables can be removed. Variable elimination repeatedly removes these variables and this way speeds up computation.

Within the example: Alarm and Earthquake are ancestor nodes of query variable JohnCalls and cannot be removed. MaryCalls is neither a query nor an evidence variable and no ancestor node. Therefore it can be removed.
Example

- Let us consider the query $P(JohnCalls \mid Burglary = true)$.

- The nested sum is

$$P(j, b) = \alpha P(b) \sum_e P(e) \sum_a P(a \mid b, e) P(j, a) \sum_m P(m \mid a)$$

- Obviously, the rightmost sum equals 1 so that it can safely be dropped.

- General observation: variables, that are not query or evidence variables and not ancestor nodes of query or evidence variables can be removed. Variable elimination repeatedly removes these variables and this way speeds up computation.
Let us consider the query $P(\text{JohnCalls} | \text{Burglary} = \text{true})$.

The nested sum is

$$P(j, b) = \alpha P(b) \sum_e P(e) \sum_a P(a | b, e) P(j, a) \sum_m P(m | a)$$

Obviously, the rightmost sum equals 1 so that it can safely be dropped.

general observation: variables, that are not query or evidence variables and not ancestor nodes of query or evidence variables can be removed. Variable elimination repeatedly removes these variables and this way speeds up computation.

within example: Alarm and Earthquake are ancestor nodes of query variable JohnCalls and cannot be removed. MaryCalls is neither a query nor an evidence variable and no ancestor node. Therefore it can be removed.
If the network is singly connected or a polytree (at most one undirected path between two nodes in the graph), the time and space complexity of exact inference is linear in the size of the network.

The burglary example is a typical singly connected network.

For multiply connected networks inference in Bayesian Networks is NP-hard.

There are approximate inference methods for multiply connected networks such as sampling techniques or Markov chain Monte Carlo.
Lecture Overview

1. Motivation
2. Foundations of Probability Theory
3. Probabilistic Inference
4. Bayesian Networks
5. Alternative Approaches
Other Approaches (1)

- Rule-based methods with "certainty factors".
  - Logic-based systems with weights attached to rules, which are combined using inference.
  - Had to be designed carefully to avoid undesirable interactions between different rules.
  - Might deliver incorrect results through overcounting of evidence.
  - Their use is no longer recommended.
Other Approaches (2)

- **Dempster-Shafer Theory**

  - Allows the representation of *ignorance* as well as uncertainly.
  
  - Example: If a coin is fair, we assume $P(Heads) = 0.5$. But what if we do not know if the coin is fair? $\rightarrow \ Bel(Heads) = 0$, $Bel(Tails) = 0$.
  
  - If the coin is 90% fair, $0.5 \times 0.9$, i.e. $Bel(Heads) = 0.45$.
  
  $\rightarrow$ Interval of probabilities is $[0.45, 0.55]$ with the evidence, $[0, 1]$ without.
  
  $\rightarrow$ The notion of utility is not yet well understood in Dempster-Shafer Theory.
Other Approaches (3)

- **Fuzzy logic and fuzzy sets**
  - A means of representing and working with *vagueness*, not uncertainty.
  - Example: The car is *fast*.
  - Used especially in control and regulation systems.
  - In such systems, it can be interpreted as an *interpolation technique*. 
Bayesian Networks allow a compact representation of joint probability distribution.

Bayesian Networks provide a concise way to represent conditional independence in a domain.

Inference in Bayesian networks means computing the probability distribution of a set of query variables, given a set of evidence variables.

Exact inference algorithms such as variable elimination are efficient for poly-trees.

In complexity of belief network inference depends on the network structure.

In general, Bayesian network inference is NP-hard.