Uncertainty

Modelling

Introduction to Artificial Intelligence

Chandra Gummaluru
University of Toronto

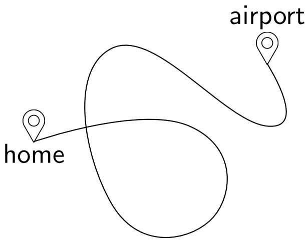
Version W22.1

- The following is based on material developed by many individuals, including (but not limited to):
    - Sheila McIlraith
    - Bahar Aameri
    - Fahiem Bacchus
    - Sonya Allin

- Part of being an intelligent agent involves being able to make decisions even in uncertain situations.

Example: Catching a Flight

- You are running late for your flight and can either:
  - Take the train
  - Take a car
- Since so many factors influence the travel time, there is no obvious choice.
- We must still make a decision, but we want to make it as rationally as possible.



- In every case we considered so far, our actions have deterministic consequences.
- We now consider relaxing this assumption, but to keep it simple, we will once again assume we are the only player.
- In other words, the new assumptions are:
  - there is only one player
  - the actions are non-deterministic, i.e., $S(s, a)$, for any $a \in A(s)$ is a random variable.

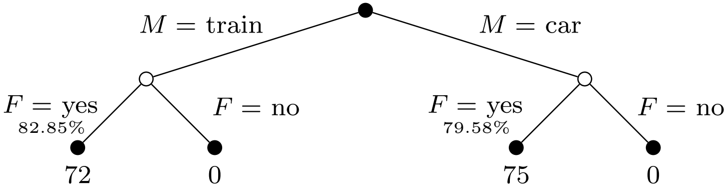
- **Example:** Catching a Flight
    - You are running late for your flight and can either:
        - Take the train
        - Take a car
    - Let $M \in \{\text{train}, \text{car}\}$ and $F \in \{\text{yes}, \text{no}\}$ be the method of transport we choose and whether we catch the flight or not.
    - Obviously, we want to catch the flight, but might have a preference in how we get there. For example, we might have

$$r(M) = \begin{cases} 6, M = \text{train} \\ 8, M = \text{car} \end{cases}$$

    - Suppose we knew that

$$P(F = \text{yes}|M) = \begin{cases} 0.8285, M = \text{train} \\ 0.7958, M = \text{car} \end{cases}$$

- We can model this situation as a 2-player game in which we must choose between $M =$ train and $M =$ car, and the adversary (representing chance) then decides whether $F =$ yes or $F =$ no according to $P(F = \text{yes}|M)$.



- We can then decide on the action using the Expect-Max strategy.
- Our expected utility is approximately 5.9652 if we choose to take the train, and 5.9685 if we choose to take the car; thus we should take the car.
- Later, we will see where we got the values for $P(F = \text{yes}|M)$.

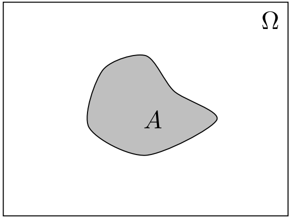
- Probability theory seeks to quantify uncertainty.
- In this context, an uncertain situation is called an **experiment**.
- The **sample space** of an experiment, denoted Ω, is the set of all outcomes:
  - When the experiment is conducted, exactly one outcome, $\omega \in \Omega$ will occur.
- In general, Ω can be discrete or continuous, but we will focus only on the former.

- The **probability** of an event, $\omega \in \Omega$, is the limit of its relative frequency as the experiment is repeated infinitely many times, i.e.,

$$\Pr\{\omega\} = \lim_{N \to \infty} \frac{N(\omega)}{N},$$

where $N(\omega)$ denotes the number of times that ω occurs.

- An **event**, A, is a subset of the outcomes, i.e., $A \subseteq \Omega$.
- We say that A occurs iff some outcome, $\omega \in A$ occurs.



- The probability of A is the sum of the probabilities of the outcomes that make it up:

$$\Pr\{A\} = \sum_{\omega \in A} \Pr\{\omega\}.$$

- Two events can share outcomes, and hence, can occur simultaneously.

- Actually repeating an experiment infinitely many times to calculate probabilities is not practical. Thus, we must manually assign them.

- A **probability measure** is a function, P, such that for each event, A,

$$\Pr\{A\} = P(A).$$

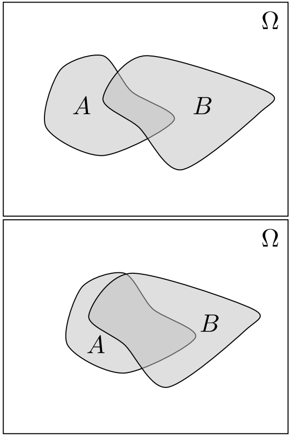
- The exact mapping is irrelevant, but it must satisfy a few axioms. In particular, $P : \mathcal{P}(\Omega) \to \mathbb{R}$ is a probability measure if and only if it is

  1. **non-negative**: $P(A) \geq 0$, for any event, A
  2. **additive**: $P(A \cup B) = P(A) + P(B)$ if $A \cap B = \emptyset$
  3. **normalized**: $P(\Omega) = 1, P(\emptyset) = 0$
  4. **complimentary**: $P(\neg A) = 1 - P(A)$

- The **joint** probability of two events, A, and B, is the probability that they both occur simultaneously, i.e.,

$$\Pr\{A \text{ and } B\} = P(A \cap B).$$

- In general, we cannot determine the joint probability from $P(A)$ and $P(B)$:
  - We can modify A and B such that $P(A \cap B)$ is different, but $P(A)$ and $P(B)$ remain unchanged.
- Thus, we must manually assign joint probabilities.



- So far, we have computed the probability of an event, A, assuming that any outcome in the sample space can occur.
- However, in many cases, we know that the outcome that occurs must be from a subset of the sample space, say B.
- Thus, as we repeat the experiment infinitely many times to compute the probabilities, we ignore the trials in which B does not occur.
- In other words, we define the **conditional** probability of A given B as:

$$\Pr\{A \text{ given } B\} = \lim_{N \to \infty} \frac{N(A \cap B)}{N(B)}.$$

- Notionally, we will write $\Pr(A \text{ given } B) = P(A|B)$.
- We can show that $P(A|B)$ can be expressed in terms of $P(A \cap B)$ and $P(B)$.
- Indeed, we have

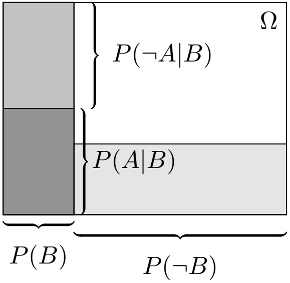
$$
\begin{aligned}
P(A|B) &= \lim_{N\to\infty} \frac{N(A \cap B)}{N(B)} \\
&= \lim_{N\to\infty} \frac{N(A \cap B)}{N} \frac{N}{N(B)} \\
&= \lim_{N\to\infty} \frac{N(A \cap B)}{N} \lim_{N\to\infty} \frac{N}{N(B)} \\
&= \frac{P(A \cap B)}{P(B)}, P(B) \neq 0.
\end{aligned}
$$

- Since $A \cap B = B \cap A$, it follows that

$$P(B \cap A) = P(A \cap B)$$

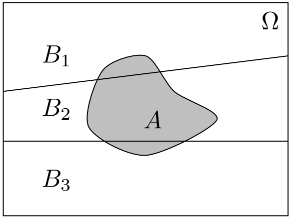
$$\Leftrightarrow P(A)P(B|A) = P(B)P(A|B)$$

$$\Leftrightarrow \underbrace{P(B|A)}_{\substack{\text{posterior of} \\ B \text{ given } A}} = \underbrace{P(B)}_{\substack{\text{prior} \\ \text{of } B}} \underbrace{\frac{P(A|B)}{P(A)}}_{\text{liklihood}}$$



- This is called **Bayes' rule**; it lets us update our beliefs given new knowledge.

- A **partition** of Ω, is a disjoint set of events, $B_1, \ldots, B_n$ such that
$$\bigcup_{i=1}^{n} B_i = \Omega.$$
- Given a partition, $B_1, \ldots, B_n$, we can write
$$P(A) = \sum_{i=1}^{n} P(A \cap B_i),$$
for any event, A.
- In some cases, it is easiest to compute $P(A)$ through such a partition.



- Two events, A and B are **independent**, denoted $A \perp B$ (or $B \perp A$) iff

$$P(B|A) = P(B) \text{ or } P(A|B) = P(A).$$

- In other words, for independent events, knowledge of one does not influence the probability of the other.
- Two events, A and B, are **conditionally independent** given another event, C, denoted $A \perp B|C$ (or $B \perp A|C$) iff

$$P(B|A, C) = P(B|C) \text{ or } P(A|B, C) = P(A|C).$$

- In other words, for conditionally independent events, knowledge of one does not influence the probability of the other, provided the condition is satisfied.
- There is no relationship between independence and conditional independence.

- To reason about events more mathematically, it is useful to map them to a numerical space.

- Such a mapping is called a **random variable**.

- Since we are assuming Ω is discrete, we can simply use $\mathbb{N}$ as the numerical space.

- For any (discrete) random variable, $X : \Omega \to \mathbb{N}$, we define a **probability mass function**, $p_X : \mathbb{N} \to [0, 1]$ such that
$$p_X(x) := \Pr\{X = x\} = P(\{\omega \in \Omega : X(\omega) = x\}).$$

- We can similarly define a **joint probability mass function**, $p_{X,Y} : \mathbb{N}^2 \to [0, 1]$ between two random variables, X, Y such that
$$p_{X,Y}(x, y) := \Pr\{X = x \text{ and } Y = y\} = P(\{\omega \in \Omega : X(\omega) = x, Y(\omega) = y\}).$$

- The joint probability mass function, $p_{X,Y}$, provides complete information about the behaviour of X and Y.

- Indeed we can compute:
  - the **conditional probability mass function**, $p_{X|Y} : \mathbb{N}^2 \to [0,1]$ of X given Y, as

$$p_{X|Y}(x,y) := \frac{p_{X,Y}(x,y)}{p_Y(y)}, p_Y(y) \neq 0.$$

  - the (marginal) probability mass function, p_X (or p_Y analogously) as:

$$p_X(x) = \Pr\{X = x\}$$

$$= \sum_{\forall y} \Pr\{X = x \text{ and } Y = y\}$$

$$= \sum_{\forall y} p_{X,Y}(x,y).$$

- An uncertain situation can be modelled as a set of random, but related variables.

- In general, these relationships can be incredibly complex.

  Example: Catching a Flight
  - Whether we catch the flight or not depends on a multitude of factors, such as:
    - when we get to the airport, and
    - how long it takes to get through security.
  - Each of the aforementioned factors, may themselves depend on other factors such as:
    - how we get to the airport, or
    - the number of bags we have.
  - However, these factors influence the outcome in different ways:
    - how we get to the airport only matters if we do not know when we get there.
    - the number of bags matters even if we know long it takes to get through security since the former will influence when we get to the airport.

- We need a systematic way to represent dependence relationships.

- A **Bayesian network** is a directed acyclic graph, $(\mathcal{V}, \mathcal{E})$, in which the vertices represent the related random variables, and the edges represent the dependence relationships between them:
    - If $(V_1, V_2) \in \mathcal{E}$, where $V_1, V_2 \in \mathcal{V}$, then V_1 and V_2 are dependent.
    - However, the converse is not necessarily true.
    - The directions of the arcs is technically irrelevant since dependence is commutative.
    - The arcs typically indicate the direction of causality but need not be.

- **Example**: Causality in a Bayesian Network
    - Both networks below indicate that X_1 and X_2 are dependent, but the left suggests that X_1 causes X_2, while the right suggests that X_2 causes X_1.

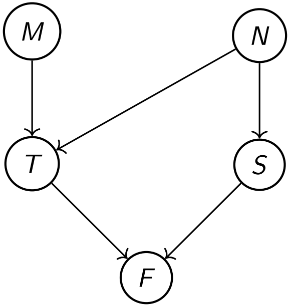
$$X_1 \longrightarrow X_2 \qquad\qquad X_1 \longleftarrow X_2$$

Example: Catching a Flight

- Define the following variables:
  - whether we catch the flight or not, F
  - when we get to the airport, T
  - how long it takes to get through security, S
  - how we get to the airport, M
  - how many bags we have, N
- We can define a Bayesian network over

$$\mathcal{V} = \{F, T, S, M, N\}.$$

- One possible network is shown on the right.
- The network says the following:
  - M influences T
  - N influences T and S
  - T and S influence F



- When expressed as a Bayesian network, we can graphically determine the conditional dependence relationships between any subset of the variables given knowledge of another subset of the variables.

- To develop it, we shall first consider the so-called "junction" network.

- A **junction**, $\mathcal{J} = (\{X_1, X_2, X_3\}, \{(e_1, e_2)\})$ is a Bayesian network that consists of three random variables, X_1, X_2, X_3, connected by two arcs, e_1 and e_2.

- It turns out that conditional mutual dependence relationships between the variables are directly tied to their causal relationships.

- Three types of causal relationships are possible.

Junction Networks: Causal Chain

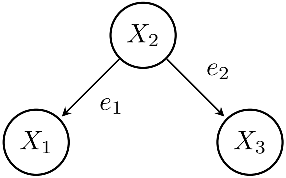
1. When one variable influences another, which in turn influences the third, as shown below, the relation is called a **causal chain**:

$$X_1 \xrightarrow{e_1} X_2 \xrightarrow{e_2} X_3$$

E.g: Rain causes the grass to get wet, which then causes worms to come out.

In this case, $X_1 \not\perp X_3$ but $X_1 \perp X_3 | X_2$.

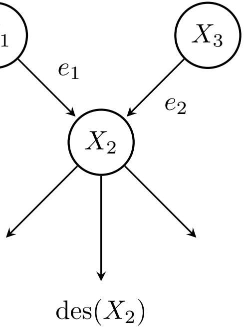
2. When one variable directly influence the other two, as shown below, the relation is referred to as a **common cause**:



E.g: Rain causes your hair to get wet and the road to be slippery.

In this case, $X_1 \not\perp X_3$ but $X_1 \perp X_3 | X_2$.

3. When one variable is directly influenced by the other two, as shown below, the relation is called a **common effect**:



E.g: Both rain and watering the garden can cause the grass to be wet.

In this case, $X_1 \perp X_3$, but $X_1 \not\perp X_3 | K$, where $K \subseteq X_2 \cup \mathrm{des}(X_2)$ and $K \neq \emptyset$.

- Let $\mathcal{B} = (\mathcal{V}, \mathcal{E})$ be some Bayesian network and $\mathcal{K} \subseteq \mathcal{V}$.
- A junction, $\mathcal{J} = (\{X_1, X_2, X_3\}, \{(e_1, e_2)\})$, of $\mathcal{B}$, is **blocked** under K if X_1 and X_2 are independent given K.
- Let a path, p, in $\mathcal{B}$ be an ordered list of junctions,

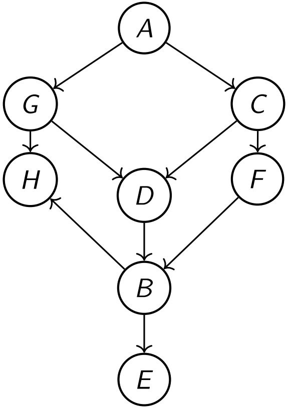
$$\mathcal{J}^{(i)} = \left(\left\{X_1^{(i)}, X_2^{(i)}, X_3^{(i)}\right\}, \left\{(e_1^{(i)}, e_2^{(i)})\right\}\right)$$

  such that $X_1^{(i+1)} = X_2^{(i)}$ and $X_2^{(i+1)} = X_3^{(i)}$, or equivalently, $e_1^{(i+1)} = e_2^{(i)}$.
- The path, p, is **blocked** under K if $\mathcal{J}^{(i)}$ is blocked under K for some i.
- **Theorem**: Conditional Independence in Bayesian Networks
  - The variables, $V_1, V_2 \in \mathcal{V}$ are conditionally independent given K if and only if every path between V_1 and V_2 is blocked under K.

Example: Dependence Separation

- We wish to determine if A and E are independent in each of the following cases:

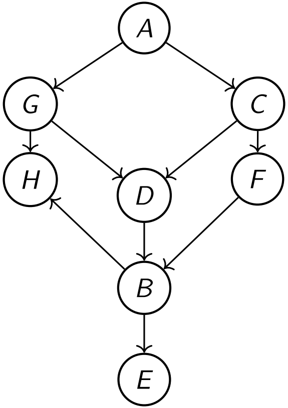
evidence	$A \perp E$
$\emptyset$	
$\{C\}$	
$\{G, C\}$	
$\{G, C, H\}$	
$\{G, F\}$	
$\{F, D\}$	
$\{F, D, H\}$	
$\{B\}$	
$\{H, B\}$	
$\{G, C, D, H, F, B\}$	



Example: Dependence Separation

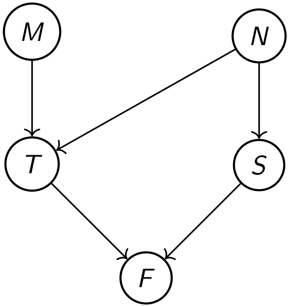
- We wish to determine if A and E are independent in each of the following cases:

evidence	$A \perp E$
$\emptyset$	no
$\{C\}$	no
$\{G, C\}$	yes
$\{G, C, H\}$	yes
$\{G, F\}$	no
$\{F, D\}$	yes
$\{F, D, H\}$	no
$\{B\}$	yes
$\{H, B\}$	yes
$\{G, C, D, H, F, B\}$	yes



Example: Catching a Flight (Dependence Separation of Variables)

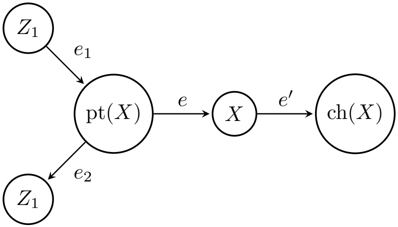
- If we are given T, then F is independent of M, but dependent on N.
- If we are also given S, then F is independent of both M and N.



- In general, we can show that any variable X, in a Bayesian network is independent of its non-descendants, given its parents,i.e.,

$$P(X|S \cup \mathrm{pts}(X)) = P(X|\mathrm{pts}(X)).$$

- To see this, we must first observe that every path from X to any of its non-descendants must contain either:
  - a causal chain or common cause junction through one of X's parents, or
  - a common effect junction through one of X's children.



- Thus, all paths to X from any of its non-descendants are blocked given $\text{pts}(X)$.
- Therefore, X is independent of its non-descendants given its parents, i.e.,

$$P(X|S \cup \text{pts}(X)) = P(X|\text{pts}(X)),$$

where S is a subset of X's non-descendants.

- This is a very important property that we will exploit later on when we develop inference techniques on Bayesian networks.