Artificial Intelligence

academic year 2024/2025

Giorgio Fumera

Pattern Recognition and Applications Lab
Department of Electrical and Electronic Engineering
University of Cagliari (Italy)

Knowledge Representation and Inference under Uncertainty

Suggested textbooks

This course requires a basic knowledge of probability theory.

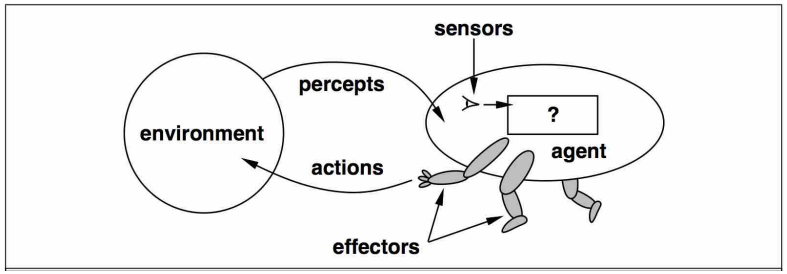
Bayesian neworks are covered by the course textbook, which provides also an informal introduction to probability theory:
S. Russell, P. Norvig, *Artificial Intelligence – A Modern Approach*, 4th Ed., Pearson, 2021 (or a previous edition)

For a formal and comprehensive introduction to probability theory, textbooks like the following one are suggested: A.M. Mood, F.A. Graybill, D.C. Boes, *Introduction to the Theory of Statistics*, McGraw-Hill, 1991 / 1998

Introduction

Introduction

Often rational agents must make decisions and act under **uncertainty** about their environments, e.g.:

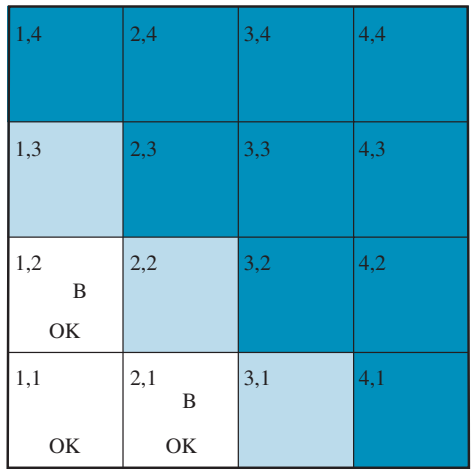


- the wumpus world (sensors report only local information)
- medical diagnosis from patients' symptoms and outcomes of medical tests
- speech/image recognition from (noisy) audio/video signals
- self-driving vehicles (incomplete and noisy sensory data)

Limitations of logical agents

Logical agents can only deal with propositions that are either true, false, or unknown, but they cannot represent a **degree of belief** about propositions.

Example: consider the agent's knowledge about the configuration of the wumpus world show on the right, after having explored squares $(1,1), (1,2)$ and $(2,1)$.



A logical agent cannot conclude anything about which of the reachable squares $(1,3), (2,2)$ and $(3,1)$ is **most likely** to be safe from pits: therefore it can only decide the next move **randomly**.

Limitations of logical agents

Diagnosis problems (e.g., in medicine) are a good example of the limitations of the logical approach in dealing with uncertainty.

Example: the following dental diagnosis rule (where p denotes a patient) is wrong, since toothache can have other causes besides cavity:

$$\forall p \; Symptom(p, Toothache) \Rightarrow Disease(p, Cavity)$$

Trying to specify **all** possible causes is pointless, e.g.:

$$\forall p \; Symptom(p, Toothache) \Rightarrow$$
$$Disease(p, Cavity) \vee Disease(p, Abscess) \vee \ldots$$

Turning the first rule into a **causal** one does not work, either, since not all cavities cause toothache:

$$\forall p \; Disease(p, Cavity) \Rightarrow Symptom(p, Toothache)$$

Making rational decisions under uncertainty

Uncertainty can be due to **laziness**, or by **theoretical** or **practical ignorance**.

In domains involving uncertainty (medicine, business, law, etc.), including AI, only a **degree of belief** in sentences of interest can be provided.

Probability theory is a widely used tool to summarise this kind of uncertainty into a numerical value, conventionally in the range $[0, 1]$.

For instance, stating that the probability that a patient with a toothache has a cavity is 0.8 means that

- in all possible situations indistinguishable by the agent from the current one, 80% of patients suffering from toothache have a cavity
- the missing 20% summarises **all** the other possible causes of toothache the agent is too lazy or ignorant to confirm or deny

Making rational decisions under uncertainty

An effective decision-making agent under uncertainty should have **preferences** about the possible **outcomes** of its actions.

Representing and reasoning with preferences is the subject of **utility theory**.

Rational decision-making under uncertainty can be achieved by **combining** the likelihood (probability) and the preference of actions' outcomes, which in turn is the subject of **decision theory**:

$$\textit{Decision theory} = \textit{probability theory} + \textit{utility theory}$$

Basically, a **rational agent** chooses the actions that yields the highest **expected** utility, averaged over all its possible outcomes.

Elements of probability theory

A notation for probability theory

In the following, a version of probability theory suited to AI applications will be presented, as a **formal language** for knowledge representation and reasoning under uncertainty.

This formal language is characterised by two elements:

- the kind of **sentence** to which degrees of beliefs are assigned
- the distiction between **prior** and **conditional** probability statements, related to the **evidence** available to the agent

The considered notation is an extension of **propositional logic**.

Propositions and random variables

Degrees of belief are assigned to **propositions** (natural language statements that can be either true or false), that refer to the "state of the world" of interest to the agent.

Propositions used in probability theory describe the "state of the world" in terms of a predefined set of **random variables**, each one referring to a "part" of the world.

Each random variable has a **domain**, i.e., the set of **mutually exclusive** values it can take.

Conventionally, random variables are represented by uppercase names (e.g., *Weather*) and their values as lowercase names (e.g., *sunny*).

The domain of random variables

The domain of random variables can be of three kinds:

- **Boolean**: $\{true,\ false\}$
- **discrete**: a set of countable values
  - categorical, e.g.: $\{sunny,\ rainy,\ cloudy,\ snow\}$ for a random variable *Weather*
  - numerical, e.g., integer values
  - Boolean (a particular case of categorical values)
- **continuous**: real numbers, e.g., the interval $[0, 1]$

This course will consider only **discrete** random variables.

It turns out that also for unordered categorical domains it is convenient to impose an (arbitrary) **ordering** on their values: therefore categorical domains will be written as **tuples**, e.g., $\langle true,\ false\rangle$, instead of sets.

Elementary and complex propositions

Example: two Boolean random variables related to a given person

- *Cavity*: whether the lower left wisdom tooth has a cavity
- *Toothache*: whether that person suffers from toothache

Elementary propositions assert that a **single** random variable has a particular value from its domain, e.g.:

$$Cavity = true$$

Complex propositions combine elementary ones using standard logical connectives, e.g.:

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

Each proposition can be assigned a **degree of belief**.

Atomic events

Atomic event: a complete description of the "state of the world" about which the agent is uncertain.

It is defined by a proposition that assigns values to **all** the random variables used by the agent.

Example: an agent uses only the two Boolean random variables *Cavity* and *Toothache*

- there are **four** distinct atomic events
- one of them is:

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

Properties of atomic events

- Atomic events are **mutually esclusive** and **exhaustive**: exactly one of them must be the case
- Any atomic event entails the truth or falsehood of every proposition; e.g., $Cavity = true \ \wedge \ Toothache = false$ entails
  - the truth of $Cavity = true$
  - the falsehood of $Cavity \Rightarrow Toothache$
- Any proposition is logically equivalent to the disjunction of all atomic events that entail its truth, e.g.:

$$\begin{aligned} Cavity = true \quad \equiv \quad & (Cavity = true \wedge Toothache = true) \ \vee \\ & (Cavity = true \wedge Toothache = false) \end{aligned}$$

Prior probability

The degree of belief p assigned to a given proposition $\mathcal{A}$, *in the absence of any other information*, is called **unconditional** or **prior probability**.

The usual notation is $P(\mathcal{A}) = p$. Some examples:

$$P(\mathit{Weather} = \mathit{sunny}) = 0.35$$
$$P(\mathit{Cavity} = \mathit{true} \wedge \mathit{Toothache} = \mathit{true}) = 0.15$$

Probability distribution

Prior probability distribution: the assignment of prior probabilities to **all** the values in the domain of a discrete random variable.

Example: a possible probability distribution for *Weather*, having domain $\langle sunny, rain, cloudy, snow \rangle$:

$$P(Weather = sunny) = 0.7$$
$$P(Weather = rain) = 0.2$$
$$P(Weather = cloudy) = 0.08$$
$$P(Weather = snow) = 0.02$$

In vector notation:

$$\mathbf{P}(Wheater) = \langle 0.7, 0.2, 0.08, 0.02 \rangle$$

The axioms of probability

The foundations of probability theory are defined as a set of axioms. In particular, **Kolmogorov's axioms**:

1. $0 \leq P(\mathcal{A}) \leq 1$, for any proposition $\mathcal{A}$
2. $P(\mathit{true}) = 1, \;\; P(\mathit{false}) = 0$
   - *true* corresponds to the occurrence of **any** atomic event, which is a **certain** event
   - *false* corresponds to the occurrence of **no** atomic event, which is an **impossible** event
3. $P(\mathcal{A} \vee \mathcal{B}) = P(\mathcal{A}) + P(\mathcal{B}) - P(\mathcal{A} \wedge \mathcal{B})$, for any pair of propositions $\mathcal{A}$ and $\mathcal{B}$

Consequences of the axioms of probability

From the axioms of probability many **theorems** can be derived.

Examples:

- for any event $\mathcal{A}$: $P(\mathcal{A}) + P(\neg\mathcal{A}) = 1$
- the values of the probability distribution of a discrete random variable X with domain $\langle x_1, x_2, \ldots, x_n \rangle$ must sum to 1 (this is a particular case of the previous theorem):

$$\sum_{i=1}^{n} P(X = x_i) = 1$$

- the probability of a proposition $\mathcal{A}$ equals the sum of the probabilities of all the atomic events $\mathcal{E}_k$ in which it holds, i.e., if $\mathcal{A} = \cup_k \mathcal{E}_k$, then:

$$P(\mathcal{A}) = \sum_{k} P(\mathcal{E}_k)$$

Joint probability distribution

It is often useful to consider the probability of the **conjunction** of values of several random variables, e.g.:

$$P(Cavity = true \wedge Weather = sunny)$$

The probabilities of **all** the combinations of values of a set of random variables is called **joint probability distribution**.

Example: the joint probability distribution $\mathbf{P}(Cavity, Weather)$ can be represented by a 4×2 table of probabilities:

		Weather			
		sunny	*rain*	*cloudy*	*snow*
Cavity	*true*	0.15	0.10	0.03	0.01
	false	0.55	0.10	0.05	0.01

Note that, according to the axioms of probability, the probabilities of a joint distribution sum to 1.

Full joint probability distribution

The joint probability distribution of **all** the random variables used by an agent is called **full joint probability distribution**.

Example: if an agent uses only the variables *Cavity*, *Toothache* and *Weather*, the full joint distribution is

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

which can be represented by a $2 \times 2 \times 4$ table of 16 probabilities.

Note that the full joint distribution specifies the probability of **every** *atomic* event.

Conditional probability

If **evidence** about some random variables has been obtained, **conditional** or **posterior** probabilities of the remaining random variables have to be considered.

Example: a patient is **known** to suffer from toothache, and **no other** information is yet available

- the **prior** probability $P(Cavity = true)$ does not reflect anymore the current state of knowledge
- the **conditional** probability of the patient's having a cavity **given that** she suffers from toothache should be considered, e.g., in the conventional notation:

$$P(Cavity = true | Toothache = true)$$

Conditional probability

Conditional probabilities are **defined** in terms of unconditional ones. For any pair of propositions $\mathcal{A}$ and $\mathcal{B}$:

$$P(\mathcal{A}|\mathcal{B}) \triangleq \frac{P(\mathcal{A} \wedge \mathcal{B})}{P(\mathcal{B})}$$

As a particular case, for any pair of random variables X and Y and any pair of values x and y in the respective domains:

$$P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}$$

Example:

$$P(\mathit{Cavity} = \mathit{true}|\mathit{Toothache} = \mathit{true}) = \frac{P(\mathit{Cavity} = \mathit{true}, \mathit{Toothache} = \mathit{true})}{P(\mathit{Toothache} = \mathit{true})}$$

Conditional probability distribution

For any pair of random variables X and Y and values y in the domain of Y, the **conditional probability distribution** of X given $Y = y$ can be written in vector notation:

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

Example: $\mathbf{P}(\mathit{Cavity}|\mathit{Toothache} = \mathit{true}) = \langle 0.8, 0.2 \rangle$

Note that:

- also the values of a conditional distribution sum to 1
- **every** value of the conditioning variable corresponds to a **distinct** conditional distribution, e.g.:
  - $\mathbf{P}(\mathit{Cavity}|\mathit{Toothache} = \mathit{true}) = \langle 0.8, 0.2 \rangle$
  - $\mathbf{P}(\mathit{Cavity}|\mathit{Toothache} = \mathit{false}) = \langle 0.05, 0.95 \rangle$

Product rule

From the **definition** of conditional probability the so called **product rule** immediately follows:

$$P(\mathcal{A} \wedge \mathcal{B}) = P(\mathcal{A}|\mathcal{B})P(\mathcal{B})$$

As a particular case, for any pair of random variables X and Y and any pair of values x and y in the respective domains:

$$P(X = x, Y = y) = P(X = x|Y = y)P(Y = y)$$

Their joint distribution can now be rewritten in vector notation:

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

where the right-hand side denotes element-wise multiplication of the corresponding table values, **not** matrix multiplication.

Marginal probability distribution

Given the joint distribution of any set of random variables $\mathbf{P}(X_1, \ldots, X_n)$, the prior distribution of any subset of them, e.g., $\mathbf{P}(X_1, \ldots, X_m)$, with $1 \leq m < n$, is called **marginal probability distribution**.

Example: Consider the joint distribution $\mathbf{P}(\mathit{Cavity}, \mathit{Weather})$ shown before:

		Weather			
		sunny	*rain*	*cloudy*	*snow*
Cavity	*true*	0.15	0.10	0.03	0.01
	false	0.55	0.10	0.05	0.01

$\mathbf{P}(\mathit{Cavity})$ and $\mathbf{P}(\mathit{Weather})$ are marginal distributions with respect to $\mathbf{P}(\mathit{Cavity}, \mathit{Weather})$.

Marginal probability distribution

How can one compute a marginal distribution from a **known** joint distribution, e.g., $\mathbf{P}(Cavity)$ from $\mathbf{P}(Cavity, Weather)$?

To this aim, a result mentioned above can be exploited, i.e.: *the probability of any proposition equals the sum of the probabilities of all the atomic events in which it holds*.

For instance, $P(Cavity = true)$ is the sum of all the values of $\mathbf{P}(Cavity, Weather)$ where $Cavity = true$:

$$P(Cavity = true) = 0.15 + 0.10 + 0.03 + 0.01 = 0.39$$

Marginalisation, or sum rule

In general, for any **disjoint** sets of variables $\mathbf{Y}$ and $\mathbf{Z}$:

$$\mathbf{P}(\mathbf{Y}) = \sum_{\mathbf{z}} \mathbf{P}(\mathbf{Y}, \mathbf{Z} = \mathbf{z}) \ ,$$

where the sum is over **all** combinations of values of the variables $\mathbf{Z}$.

This process is called **marginalisation**, or **sum rule**, and can be applied if the joint distribution $\mathbf{P}(\mathbf{Y}, \mathbf{Z})$ is **known**.

It is often useful to rewrite the sum rule in terms of the **product** rule, using **conditional** probabilities:

$$\mathbf{P}(\mathbf{Y}) = \sum_{\mathbf{z}} \mathbf{P}(\mathbf{Y}|\mathbf{Z} = \mathbf{z}) P(\mathbf{Z} = \mathbf{z})$$

Probabilistic inference

In most cases of interest for rational agents in AI, **probabilistic inference** consists in computing the **posterior** probability of a set of **query** variables $\mathbf{Q}$ given the **evidence** about a distinct set of variables $\mathbf{E} = \mathbf{e}$

- either for a single value of interest $\mathbf{q}$: $P(\mathbf{Q} = \mathbf{q} | \mathbf{E} = \mathbf{e})$
- or the posterior distribution $\mathbf{P}(\mathbf{Q} | \mathbf{E} = \mathbf{e})$

If the **full** joint distribution of the variables used by an agent is **known**, the **sum** and **product** rules allow the agent, in priniciple, to carry out **any** probabilistic inference, as shown in the following.

Inference using the full joint distribution: example

An agent uses three Boolean variables related to a given patient of a dentist:

- *Toothache* and *Cavity*, defined previously
- *Catch*: the dentist's steel probe catches in the lower left wisdom tooth

Assume the full joint distribution is:

	Toothache $= t$		*Toothache* $= f$	
	Catch $= t$	*Catch* $= f$	*Catch* $= t$	*Catch* $= f$
Cavity $= t$	0.108	0.012	0.072	0.008
Cavity $= f$	0.016	0.064	0.144	0.576

Inference using the full joint distribution: example

Unconditional probabilities $P(Q)$, such as

$$P(Cavity = true \vee Toothache = true) \ ,$$

can be computed as shown above, as the sum of the probabilities of the atomic events in which the event $\mathcal{Q}$ holds:

$$\begin{aligned} &P(Cavity = true \vee Toothache = true) \\ &\quad = 0.108 + 0.012 + 0.072 + 0.008 + 0.016 + 0.064 = 0.28 \end{aligned}$$

Inference using the full joint distribution: example

Conditional probabilities $P(\mathcal{Q}|\mathcal{E})$, such as

$$P(\mathit{Cavity} = \mathit{true} | \mathit{Toothache} = \mathit{true}) \ ,$$

can be computed using the definition of conditional probability, in terms of **unconditional** ones, which can in turn be computed from the full joint distribution as shown above:

$$\begin{aligned} P(\mathit{Cavity} = t | \mathit{Toothache} = t) &= \frac{P(\mathit{Cavity} = t \wedge \mathit{Toothache} = t)}{P(\mathit{Toothache} = t)} \\ &= \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064} \end{aligned}$$

Note that the last step corresponds to applying the **sum rule** to the **full** joint distribution.

Inference using the full joint distribution

A **general inference procedure** for conditional distributions $\mathbf{P}(Q|\mathbf{E} = \mathbf{e})$, for the case of a **single** query variable Q, with $\mathbf{Y}$ denoting **all** the remaining variables:

1. rewriting $\mathbf{P}(Q|\mathbf{E} = \mathbf{e})$ using the definition of conditional probability, in terms of **unconditional** probabilities
2. using the **sum rule** to compute the numerator and denominator, from the **full** joint distribution $\mathbf{P}(Q, \mathbf{E}, \mathbf{Y})$

$$\mathbf{P}(Q|\mathbf{E} = \mathbf{e}) = \frac{\mathbf{P}(Q, \mathbf{E} = \mathbf{e})}{\mathbf{P}(\mathbf{E} = \mathbf{e})} = \frac{\sum_{\mathbf{y}} \mathbf{P}(Q, \mathbf{E} = \mathbf{e}, \mathbf{Y} = \mathbf{y})}{\sum_{q,\mathbf{y}} \mathbf{P}(Q = q, \mathbf{E} = \mathbf{e}, \mathbf{Y} = \mathbf{y})}$$

Inference using the full joint distribution

To sum up, if the **full joint** distribution of the random variables of interest is known, the probability of **any** event can be computed using

- the definition of conditional probability
- the sum and product rules

In practice, two issues arise:

- **how** to compute or estimate the full joint distribution?
- what is the **computational complexity** of the inference procedure based on it?

Defining the full joint distribution

Defining the full joint distribution of variables $X_1, \ldots, X_n$ amounts to **assigning** a probability value to **each** of the **atomic events**

$$X_1 = x_1 \wedge \ldots \wedge X_n = x_n$$

To assign a probability value to **any** event, including atomic ones, three different approaches exist, based on different **definitions** of probability:

- classical
- frequentist
- subjective

Classical probability

In some cases the **atomic** events that describe the "state of the world" can be assumed to be:

- mutually exclusive
- equally likely
- random

Examples: tossing a coin, throwing a dice.

In such cases the **classical** (or **a priori**) definition probability can be applied for **any** event $\mathcal{A}$:

$$P(\mathcal{A}) = \frac{\text{number of atomic events where } \mathcal{A} \text{ holds}}{\text{total number of atomic events}}$$

Classical probability: examples

- Tossing a coin: probability of a head (or a tail)?
- Throwing a dice
  - probability of face 3 up?
  - probability of an even face up?
- Tossing **two** coins: probability of getting two heads?
- Throwing **two** dice
  - probability of getting (6,6)?
  - probability that the sum of the faces up is 6?
- Picking a card from a well shuffled deck of 52 cards
  - probability of picking an ace of hearts (♡)?
  - probability of picking an ace or a spade (♠)?

Frequentist probability

In most practical applications the assumptions of classical probabilty are **not valid**, e.g., computing $P(\mathcal{A})$ for:

- $\mathcal{A} =$ face 6 up after throwing a **loaded** dice
- $\mathcal{A} =$ a chip manufactured by company XYZ is faulty
- $\mathcal{A} =$ a MSc student in CECAI graduates within 2 years

Nevertheless, it may be possible to observe **multiple** "states of the world" under **similar** and **uniform** conditions, e.g.:

- **repeatedly** throwing a loaded dice and observing the outcomes
- testing **a sample** of the chips manufactured by company XYZ
- collecting **records** of past MSc students in CECAI

Frequentist probability

If **multiple** "states of the world" can be observed under **similar** and **uniform** conditions, the **frequentist** (or **a posteriori**) probability definition can be used.

It **estimates** the probability of **any** event $\mathcal{A}$, including atomic events, as the **relative frequency** of its occurrences:

$$P(\mathcal{A}) = \frac{\text{number of observations where } \mathcal{A} \text{ occurs}}{\text{total number of observations}}$$

Examples

- fraction of throws of a dice leading to face 6 up
- fraction of XYZ's sampled chips that are faulty
- fraction of past MSc students in CECAI who graduated within 2 years

Subjective probability

Common requirement of classical and frequentist probability: a **conceptual** "experiment" in which the outcomes can occur under **uniform conditions**.

This requirement is not always fulfilled. For instance, how to compute the probability that

- the piano player John Smith will break one or both of his hands within the next 10 years?
- the third World War will start within 2025?

In such cases **subjective** probability can be used, e.g., involving domain experts' judgement.

Computational complexity of probabilistic inference

For n discrete random variables $X_1, \ldots, X_n$ whose domains are made up of $d_1, \ldots, d_n$ values, the number of atomic events is

$$d_1 \times \ldots \times d_n$$

Therefore, the number of probability values to be assigned to define the full joint distribution (they must sum to 1) grows **exponentially** in the domain size:

$$d_1 \times \ldots \times d_n - 1$$

Example: for n Boolean variables

- the number of atomic events is 2^n
- the number of probability values to assign is $2^n - 1$

Computational complexity of probabilistic inference

Also the probabilistic inference procedure presented above has an **exponential** complexity in the domain size, due to the **sum rule**.

For instance, for n Boolean variables $X_1, \ldots, X_n$ the marginal distribution $\mathbf{P}(X_1)$ is given by

$$\sum_{x_2,\ldots,x_n \in \{0,1\}^{n-1}} \mathbf{P}(X_1, X_2 = x_2, \ldots, X_n = x_n) \ ,$$

which requires $n-2$ sums for **each** of the 2^{n-1} combinations of values $x_2, \ldots, x_n$, i.e., $\mathcal{O}(2^n)$ sums.

Therefore, the full joint distribution in tabular form is not a practical tool for building probabilistic reasoning systems.

Independence

Two events $\mathcal{A}$ and $\mathcal{B}$ are said to be **independent**, if **any** of the following **equivalent** relation holds:

$$P(\mathcal{A}|\mathcal{B}) = P(\mathcal{A}), \quad P(\mathcal{B}|\mathcal{A}) = P(\mathcal{B}), \quad P(\mathcal{B} \wedge \mathcal{A}) = P(\mathcal{A})P(\mathcal{B})$$

In particular, two variables X and Y are independent, if any of the following relation holds between their **distributions**:

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

Independence is useful to our purposes since it **reduces**

- the number of probability values to be assigned to specify the **full** joint distribution
- the computational complexity of probabilistic inference

Independence: example

Assume that an agent uses the variables

- *Toothache*, *Catch* and *Cavity* (all Boolean)
- *Wheather*, with domain $\langle sunny, rain, cloudy, snow \rangle$

The full joint distribution $\mathbf{P}(Toothache, Catch, Cavity, Wheather)$ is specified by a $2 \times 2 \times 2 \times 4$ table of 32 probability values.

Is there any way to **reduce** the number of probabilty values to assign, to specify this distribution?

To this aim, one may wonder whether and how dental problems (toothache, the dentist's probe catching in a tooth, and cavities) and the wheather are related to each other...

Independence: example

. . . Consider rewriting the full joint distribution using the **product rule**:

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

It is reasonable to assume that dental problems do not influence the wheather, i.e., the variable *Wheather* can be assumed to be **independent** from *Toothache*, *Catch* and *Cavity*:

$$\mathbf{P}(\mathit{Wheather} | \mathit{Toothache}, \mathit{Catch}, \mathit{Cavity}) = \mathbf{P}(\mathit{Wheather})$$

This allows rewriting the full joint distribution as:

$$\begin{aligned}&\mathbf{P}(\mathit{Toothache}, \mathit{Catch}, \mathit{Cavity}, \mathit{Wheather})\\ &\quad = \mathbf{P}(\mathit{Wheather})\mathbf{P}(\mathit{Toothache}, \mathit{Catch}, \mathit{Cavity})\end{aligned}$$

Independence: example

Using the last expression, the full joint distribution can be specified by:

- $\mathbf{P}(\mathit{Wheather})$: $4^1 - 1 = 3$ probability values,
- $\mathbf{P}(\mathit{Toothache}, \mathit{Catch}, \mathit{Cavity})$: $2^3 - 1 = 7$ probability values,

for a total of $3 + 7 = \mathbf{10}$ values, instead of $2^4 - 1 = \mathbf{31}$!

In the **best case**, **all** variables are independent. For instance, the full joint distribution of n Boolean variables is specified by

- $2^n - 1 = \mathcal{O}(2^n)$ values, in general (**exponential** in n)
- $n \times (2^1 - 1) = n = \mathcal{O}(n)$ values, if they are independent (**linear** in n), i.e.:

$$\mathbf{P}(X_1, \ldots, X_n) = \prod_{i=1}^{n} \mathbf{P}(X_i)$$

Unfortunately, in practice **absolute** independence is quite rare, even among subsets of random variables.

Conditional independence

Consider the joint distribution $\mathbf{P}(\mathit{Toothache}, \mathit{Catch}, \mathit{Cavity})$, which is specified by $2^3 - 1 = 7$ values.

These variables, or subsets of them, **cannot** be assumed to be independent. For instance, if the probe catches in the tooth, probably there is a cavity, and that probably causes a toothache.

Noting that cavity is a possible **cause** of **effects** like toothache and of the dentist's probe catching in a tooth, let us rewrite the joint distribution using the **product rule** in the form $P(\mathit{effect}|\mathit{cause})$:

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

What can we say about $\mathbf{P}(\mathit{Toothache}, \mathit{Catch}|\mathit{Cavity})$?

Conditional independence

Given the state (presence of absence) of a cavity, *Toothache* and *Catch* have no direct influence on each other, and therefore can be assumed to be **independent**, which is written as:

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

Equivalently:

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

This is an example of **conditional independence**.

This allows the joint distribution to be rewritten as

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

Conditional independence

How many probability values must be assigned to specify the joint distribution $\mathbf{P}(\mathit{Toothache}, \mathit{Catch}, \mathit{Cavity})$, using the conditional independence assumption above?

- $\mathbf{P}(\mathit{Toothache}|\mathit{Cavity})$: $2 \times (2^1 - 1) = 2$ values,
- $\mathbf{P}(\mathit{Catch}|\mathit{Cavity})$: $2 \times (2^1 - 1) = 2$ values,
- $\mathbf{P}(\mathit{Cavity})$: $2^1 - 1 = 1$ value,

for a total of $2 + 2 + 1 = \mathbf{5}$ values, instead of $\mathbf{7}$.

This may seem a small gain...

Conditional independence

. . . However, in practical applications involving many random variables and non-Boolean domains the advantage can be notable.

For instance, consider n Boolean variables $X_1, \ldots, X_n$ conditionally independent given a Boolean variable Y:

- the full joint distribution $\mathbf{P}(X_1, \ldots, X_n, Y)$ is specified by $2^{n+1} - 1 = \mathcal{O}(\mathbf{2^{n+1}})$ probability values
- the equivalent expression $\mathbf{P}(Y)\mathbf{P}(X_1|Y) \times \ldots \times \mathbf{P}(X_n|Y)$ is specified by
  - $\mathbf{P}(Y)$: $2^1 - 1 = 1$ value,
  - $\mathbf{P}(X_i|Y), i = 1, \ldots, n$: $n \times (2^1 - 1) = n$ values,

  for a total of $n + 1 = \mathcal{O}(\mathbf{n})$ values

This means that the number of probability values to assign reduces from **exponential** to **linear** in the number of variables.

Conditional independence

In general, two variables X and Y are **conditionally independent** given a variable Z, if:

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

As with absolute independence, this is **equivalent** to:

$$\mathbf{P}(X|Y, Z) = \mathbf{P}(X|Z) \quad \text{and} \quad \mathbf{P}(Y|X, Z) = \mathbf{P}(Y|Z)$$

Conditional independence is very useful in practice:

- it allows probabilistic systems to **scale up**
- it is much more common than absolute independence

Probabilistic inference

As mentioned above, probabilistic inference in AI usually consists in computing **posterior** probabilities $P(\mathcal{Q}|\mathcal{E})$

- $\mathcal{Q}$: a **query** event
- $\mathcal{E}$: an available **evidence**

In terms of random variables, the most general form of probabilistic inference consists in computing a **conditional** probability distribution $\mathbf{P}(\mathbf{Q}|\mathbf{E})$

- $\mathbf{Q}$: a set of **query** variables
- $\mathbf{E}$: a distinct set of **evidence** variables

Probabilistic inference: Bayes' rule

Consider any pair of events $\mathcal{A}$ and $\mathcal{B}$. The probability of their conjunction can be rewritten using the product rule:

$$\begin{aligned} P(\mathcal{A} \wedge \mathcal{B}) &= P(\mathcal{A}|\mathcal{B})P(\mathcal{B}) \\ P(\mathcal{B} \wedge \mathcal{A}) &= P(\mathcal{B}|\mathcal{A})P(\mathcal{A}) \end{aligned}$$

Since $P(\mathcal{A} \wedge \mathcal{B}) = P(\mathcal{B} \wedge \mathcal{A})$, it follows that:

$$P(\mathcal{B}|\mathcal{A}) = \frac{P(\mathcal{A}|\mathcal{B})P(\mathcal{B})}{P(\mathcal{A})}$$

The above equation is known as **Bayes' rule**. It turns out to be a fundamental tool for probabilistic inference in AI systems.

Probabilistic inference: Bayes' rule

In terms of probability distributions of random variables, Bayes' rule can be written as:

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

The most general form involves **sets** of random variables:

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

Note that the denominator can be rewritten in turn as a function of the same distributions in the numerator, using the sum and product rules:

$$\mathbf{P}(\mathbf{Y}) = \sum_{\mathbf{x}} \mathbf{P}(\mathbf{Y}, \mathbf{X} = \mathbf{x}) = \sum_{\mathbf{x}} \mathbf{P}(\mathbf{Y}|\mathbf{X} = \mathbf{x})\mathbf{P}(\mathbf{X} = \mathbf{x})$$

Probabilistic inference: Bayes' rule

To compute **one** conditional probability, Bayes' rule requires **three** other probabilities. So, why is it useful in AI (and in other domains)?

The reason is that robabilistic inference is usually of **diagnostic** form:

$$P(cause|effect) \ ,$$

where the evidence consists of observed **effects** (e.g., a symptom like toothache) of a possible **cause** of interest (e.g., a cavity).

In practice, **diagnostic knowledge**, i.e., estimating $P(cause|effect)$, is more difficult to obtain than **causal knowledge**, i.e., $P(effect|cause)$.

For instance, $\mathbf{P}(Cavity|Toothache)$ can be more difficult to estimate than $\mathbf{P}(Toothache|Cavity)$.

(cont.)

Probabilistic inference: Bayes' rule

In this context, Bayes' rule allows to carry out diagnostic inference using **only** causal knowledge:

$$P(cause|effect) = \frac{P(effect|cause)P(cause)}{P(effect)}$$

Note that, as shown above, $P(effect)$ can be computed in terms of $P(effect|cause)$ and $P(cause)$: therefore, only these two probabilities need to be estimated.

Probabilistic inference using Bayes' rule: an example

A doctor knows that

- **meningitis** causes **stiff neck** in 50% of patients (**causal knowledge**)
- 1 out of 50.000 individuals in the population has meningitis
- 1 out of 20 individuals suffers from stiff neck

Using the Boolean variables M and S to denote whether a random individual has meningitis and stiff neck, respectively, the above knowledge can be formalised as:

- $P(S = true|M = true) = 0.5$
- $P(M = true) = 1/50.000$
- $P(S = true) = 1/20$

(cont.)

Probabilistic inference using Bayes' rule: an example

The doctor can be interested in computing the probability that a patient with stiff neck has meningitis (**diagnostic inference**).

She can exploit the above knowledge, using Bayes' rule:

$$P(M = true|S = true) = \frac{P(S = true|M = true)P(M = true)}{P(S = true)}$$
$$= \frac{0.5 \times 1/50.000}{1/20} = 0.0002$$

This is a particular case of the general fact mentioned above: in the medical domain, **diagnostic knowledge** (from symptoms to diseases) is usually more difficult to obtain than **causal knowledge** (from diseases to symptoms).

(cont.)

Probabilistic inference using Bayes' rule: an example

Even if diagnostic knowledge is available from statistical observation to directly estimate $P(M = true|S = true)$, what happens if there is an **epidemic** of meningitis, i.e., if $P(M = true)$ increases?

Intuitively, also $P(M = true|S = true)$ will increase, but **how to update it**?

Also in this case Bayes' rule is useful, because:

- $P(S = true|M = true)$ is **unaffected** by the epidemic
- thus $P(M = true|S = true)$ increases **proportionately** with $P(M = true)$

This is an example of how **causal knowledge** provides the necessary **robustness** to make probabilistic systems feasible in practical applications.

Using Bayes' rule for combining evidence

Consider again the variables *Toothache*, *Catch* and *Cavity*.

A dentist may collect **two** evidences: her probe chatches in the patient's aching tooth. She is therefore interested in the following conditional distribution:

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

This is a kind of **diagnostic** inference of the form

$$P(cause|effects)$$

We already know that using the full joint distribution (if known) to compute the above conditional probability does not scale up to large numbers of variables.

Using Bayes' rule for combining evidence

The doctor can try to apply Bayes' rule to turn diagnostic inference into **causal** inference of the form $P(effects|cause)$:

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

However, estimating $\mathbf{P}(Toothache, Catch|Cavity)$ still requires $2 \times (2^2 - 1) = 6$ probability values, which does not scale up to larger numbers of evidence and query variables, either.

For instance, for n evidence variables $E_1, \ldots, E_n$ and a single query variable Q, all Boolean, the conditional distributions $\mathbf{P}(E_1, \ldots, E_n|Q)$ require $2 \times (2^n - 1)$ probability values.

(cont.)

Using Bayes' rule for combining evidence

Fortunately, we have seen that a **conditional independence** assumption can be made:

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

Each distribution in the right-hand side is specified by

$$2 \times (2^1 - 1) = 2$$

probability values, for a total of 4 values instead of 6.

Using Bayes' rule for combining evidence

In general, if n Boolean evidence variables $E_1, \ldots, E_n$ are conditionally independent on a Boolean query variable Q, i.e.:

$$\mathbf{P}(E_1, \ldots, E_n | Q) = \prod_{i=1}^{n} \mathbf{P}(E_i | Q) \ ,$$

the number of probability values to specify reduces from $2 \times (2^n - 1) = \mathcal{O}(2^{n+1})$ to $n \times 2 \times (2^1 - 1) = 2n = \mathcal{O}(n)$.

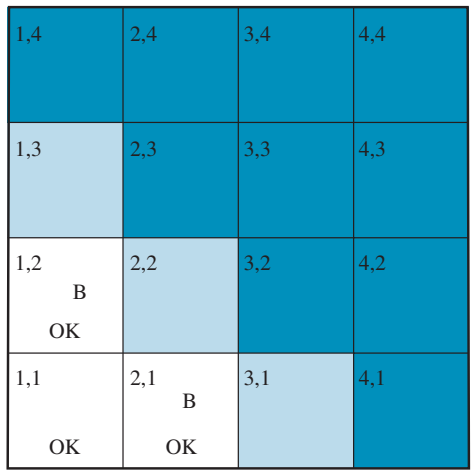
This example highlights the usefulness of **conditional independence**, together with Bayes' rule, for **diagnostic** inference of the form $P(\mathit{cause}|\mathit{effects})$ involving the combination of different pieces of evidence.

Exercise: the wumpus world revisited

Two properties of the wumpus world are that

- a pit causes breezes in all neighboring squares
- each square other than $(1, 1)$ can contain a pit with probability 0.2

Consider now the agent's knowledge about the configuration of the wumpus world shown on the right, after having explored squares $(1, 1), (1, 2)$ and $(2, 1)$.



Based on the current agent's knowledge, each of the three reachable squares $(1, 3)$, $(2, 2)$ and $(3, 1)$ can contain a pit: what is the probability that each of them contains a pit?

Bayesian networks

Probabilistic graphical models

Probabilistic graphical models: a framework to represent in graphical form the **structure** of probability distributions.

Main classes of graphical models:

- **Bayesian networks**: **directed acyclic** graphs
- **Markon random fields**: **undirected** graphs

This course focuses on Bayesian networks, which are useful for

- representing the joint distribution of random variables, expressing **causal** dependencies and **conditional independence** relations between them
- developing efficient **probabilistic inference algorithms** based on graph structure

The chain rule

Bayesian networks exploit a particular expression of the joint distribution of a set of variables based on the **product rule**.

Example: consider three variables X_1, X_2, X_3; their joint distribution $\mathbf{P}(X_1, X_2, X_3)$ can be rewritten using the product rule as:

$$\mathbf{P}(X_1, X_2, X_3) = \mathbf{P}(X_3|X_2, X_1)\mathbf{P}(X_2, X_1)$$

Applying again the product rule to $\mathbf{P}(X_2, X_1)$:

$$\mathbf{P}(X_1, X_2, X_3) = \mathbf{P}(X_3|X_2, X_1)\mathbf{P}(X_2|X_1)\mathbf{P}(X_1)$$

Different but **equivalent** expressions of the **same** joint distribution can be obtained considering a different **order** between the variables, e.g.:

$$\mathbf{P}(X_1, X_2, X_3) = \mathbf{P}(X_2|X_1, X_3)\mathbf{P}(X_1|X_3)\mathbf{P}(X_3)$$

The chain rule

In general, for a given order of n variables:

$$\begin{aligned}
&\mathbf{P}(X_1, X_2, \ldots, X_n) \\
&\quad = \mathbf{P}(X_n | X_{n-1}, \ldots, X_1)\mathbf{P}(X_{n-1} | X_{n-2}, \ldots, X_1) \cdots \mathbf{P}(X_2 | X_1)\mathbf{P}(X_1) \\
&\quad = \prod_{k=1}^{n} \mathbf{P}(X_k | X_{k-1}, \ldots, X_1)
\end{aligned}$$

This equivalence is called **chain rule**.

Note the **structure** of this expression of the joint distribution:

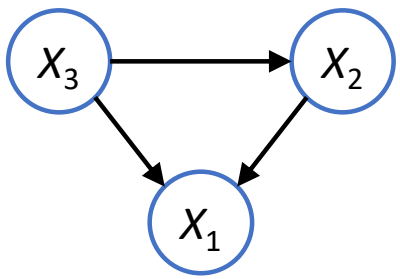
- a product of $n-1$ **conditional** and 1 **unconditional** distributions
- each term is the distribution of a **single**, distinct variable
- each conditional distribution is conditioned on all the variables that **follow** the considered one in the chosen order

Bayesian network structure

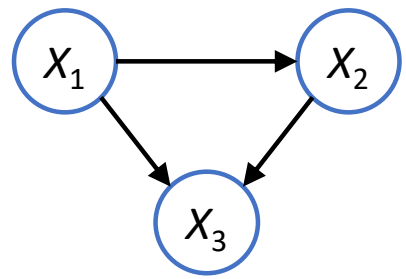
The expression of the joint distribution obtained using the **chain rule** can be represented by a **Bayesian network** (BN), which is a **directed acyclic graph** (DAG) where

- each **node** represents one of the **random variables**, and is associated with its distribution in the expression of the chain rule
- **oriented edges** represent **conditional dependencies**, linking each variable (node) with the ones on which its distribution is conditioned

Example: BNs representing two possible, **equivalent** expressions of the joint distribution $\mathbf{P}(X_1, X_2, X_3)$



$\mathbf{P}(X_1|X_2, X_3)\mathbf{P}(X_2|X_3)\mathbf{P}(X_3)$



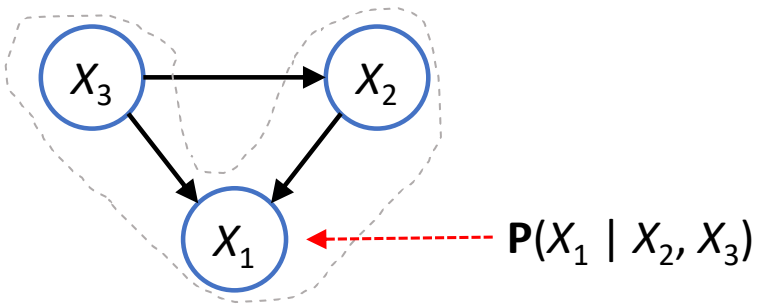
$\mathbf{P}(X_3|X_2, X_1)\mathbf{P}(X_2|X_1)\mathbf{P}(X_1)$

Bayesian network structure

In a BN the conditional distribution $\mathbf{P}(X_k|X_{k-1},\ldots,X_1)$ is represented by

- a **node** associated with X_k
- **edges** pointing **from** nodes $X_{k-1},\ldots,X_1$ **to** X_k, which are called the **parents** of X_k

Example:



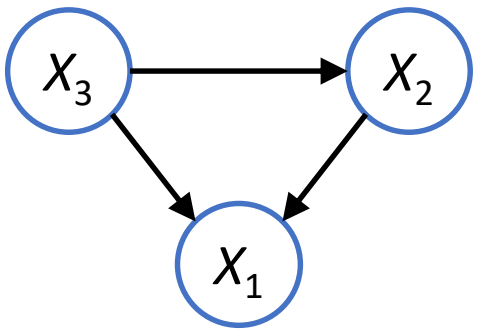
Denoting the set of parents of X_k by $pa(X_k)$, the chain rule can be concisely written as

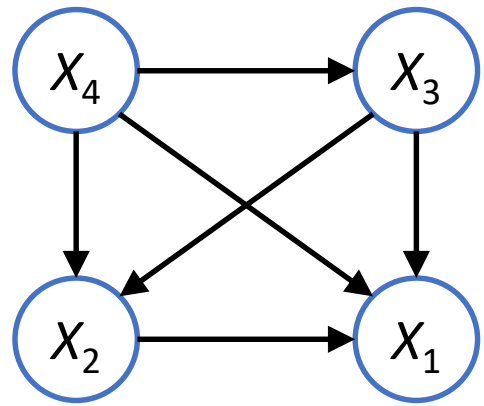
$$\mathbf{P}(X_1, X_2, \ldots, X_n) = \prod_{k=1}^{n} \mathbf{P}(X_k|pa(X_k))$$

Bayesian network structure

A BN representing a joint distribution expressed through the chain rule is a **fully connected** DAG, i.e., there is an edge (regardless of the direction) between **every** pair of nodes.

Examples:





$$\mathbf{P}(X_1, X_2, X_3) = \mathbf{P}(X_1|X_2, X_3)\mathbf{P}(X_2|X_3)\mathbf{P}(X_3)$$

$$\mathbf{P}(X_1, X_2, X_3, X_4) = \mathbf{P}(X_1|X_2, X_3, X_4)\mathbf{P}(X_2|X_3, X_4) \times \mathbf{P}(X_3|X_4)\mathbf{P}(X_4)$$

Probabilistic inference with Bayesian networks

The chain rule suggests that, instead of estimating the full joint distribution, one could estimate the corresponding conditional (or marginal) distributions, and use them for probabilistic inference.

Example: given three variables X_1, X_2, X_3, we know that any probability involving them can be computed using the product and sum rules from the full joint distribution, e.g.:

$$\mathbf{P}(X_1|X_2) = \frac{\mathbf{P}(X_1, X_2)}{\mathbf{P}(X_2)} = \frac{\sum_{x_3} \mathbf{P}(X_1, X_2, X_3 = x_3)}{\sum_{x_1, x_3} \mathbf{P}(X_1 = x_1, X_2, X_3 = x_3)}$$

Using the chain rule, the numerator could be computed, e.g., as:

$$\sum_{x_3} \mathbf{P}(X_1|X_2, X_3 = x_3)\mathbf{P}(X_2|X_3 = x_3)\mathbf{P}(X_3 = x_3)$$

Complexity of probabilistic inference with BNs

Does the chain rule provide any advantage over the full joint distribution?

If one considers the number of probability values to estimate for specifying the distributions required by the chain rule, the answer is **no**: both approaches are **equivalent** under this aspect.

Example: consider three Boolean variables, and one possible expression of the chain rule:

$$\mathbf{P}(X_1, X_2, X_3) = \mathbf{P}(X_1|X_2, X_3)\mathbf{P}(X_2|X_3)\mathbf{P}(X_3)$$

To estimate $\mathbf{P}(X_1, X_2, X_3)$ one needs to specify $2^3 - 1 = 7$ probability values.

(cont.)

Complexity of probabilistic inference with BNs

To estimate the three distributions required by the chain rule, one needs to specify:

- $\mathbf{P}(X_1|X_2, X_3)$: $2^2 \times (2^1 - 1) = 4$ values,
- $\mathbf{P}(X_2|X_3)$: $2 \times (2^1 - 1) = 2$ values,
- $\mathbf{P}(X_3)$: $2^1 - 1 = 1$ value,

for a total of 7 values, as for the full joint distribution.

Conditional indepence relations in Bayesian networks

On the other hand, we have seen that **conditional independence** can reduce the complexity of estimating conditional distributions.

Conditional independence relations can be exploited in the expression of the chain rule. For instance, consider the above example on three Boolean variables:

$$\mathbf{P}(X_1, X_2, X_3) = \mathbf{P}(X_1|X_2, X_3)\mathbf{P}(X_2|X_3)\mathbf{P}(X_3)$$

If X_1 is conditionally independent on X_2 given X_3, i.e.,

$$\mathbf{P}(X_1|X_2, X_3) = \mathbf{P}(X_1|X_3) \ ,$$

to estimate $\mathbf{P}(X_1|X_3)$ one needs to specify $2 \times (2^1 - 1) = 2$ probability values instead of $2^2 \times (2^1 - 1) = 4$.

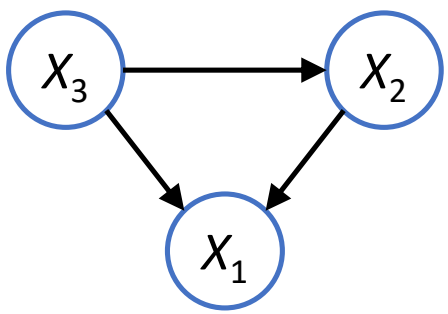
Conditional indepence relations in Bayesian networks

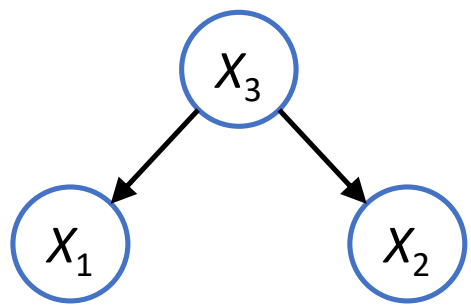
Conditional independence relations are expressed in the structure of a BN by the **absence** of the corresponding edges, leading to a DAG which is **not** fully connected.

In the previous example, the **general** expression of the chain rule corresponds to the fully connected BN shown below on the left. The expression resulting from the conditional independence assumption is:

$$\mathbf{P}(X_1, X_2, X_3) = \mathbf{P}(X_1|X_3)\mathbf{P}(X_2|X_3)\mathbf{P}(X_3) \ ,$$

which corresponds to the BN below on the right.





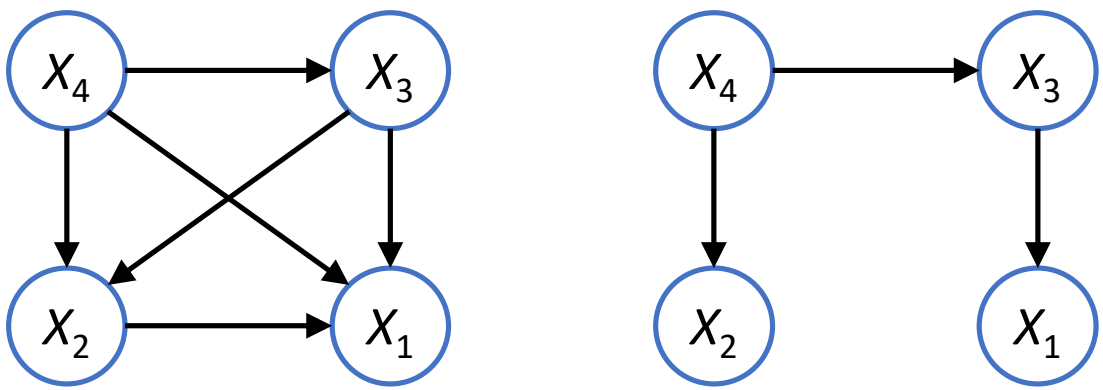
Conditional indepence relations in Bayesian networks

Example: a possible expression of the full joint distribution of four Boolean variables, corresponding to the BN shown below on the left:

$$\mathbf{P}(X_1, X_2, X_3, X_4) = \mathbf{P}(X_1|X_2, X_3, X_4)\mathbf{P}(X_2|X_3, X_4)\mathbf{P}(X_3|X_4)\mathbf{P}(X_4)$$

If X_1 is conditionally independent on X_2 and X_4 given X_3, and X_2 is conditionally independent on X_3 given X_4 (see the BN below on the right):

$$\mathbf{P}(X_1, X_2, X_3, X_4) = \mathbf{P}(X_1|X_3)\mathbf{P}(X_2|X_4)\mathbf{P}(X_3|X_4)\mathbf{P}(X_4)$$



(cont.)

Conditional indepence relations in Bayesian networks

How many proability values need to be specified to estimate the full joint distribution using the chain rule?

- General form: $\mathbf{P}(X_1|X_2, X_3, X_4)\mathbf{P}(X_2|X_3, X_4)\mathbf{P}(X_3|X_4)\mathbf{P}(X_4)$
  - $\mathbf{P}(X_1|X_2, X_3, X_4)$: $2^3 \times (2^1 - 1) = 8$ values,
  - $\mathbf{P}(X_2|X_3, X_4)$: $2^2 \times (2^1 - 1) = 4$ values,
  - $\mathbf{P}(X_3|X_4)$: $2 \times (2^1 - 1) = 2$ values,
  - $\mathbf{P}(X_4)$: $2^1 - 1 = 1$ value, for a total of **15** values
- Taking into account conditional independence relations: $\mathbf{P}(X_1|X_3)\mathbf{P}(X_2|X_4)\mathbf{P}(X_3|X_4)\mathbf{P}(X_4)$
  - $\mathbf{P}(X_1|X_3)$: $2 \times (2^1 - 1) = 2$ values,
  - $\mathbf{P}(X_2|X_4)$: $2 \times (2^1 - 1) = 2$ values,
  - $\mathbf{P}(X_3|X_4)$: $2 \times (2^1 - 1) = 2$ values,
  - $\mathbf{P}(X_4)$: $2^1 - 1 = 1$ value, for a total of **7** values

Constructing Bayesian networks

To efficiently represent full joint distributions through the chain rule and the corresponding BNs, conditional independence relations should be exploited whenever possible.

How can such relations be **identified**, and **exploited**?

Main issue: the chain rule provides $n!$ **equivalent** expressions of the full joint distribution of n variables, one for each possible **order** between them. What is the "best" order to choose?

Constructing Bayesian networks

Example: two possible expressions of the chain rule for the full joint distribution of four variables, $\mathbf{P}(X_1, X_2, X_3, X_4)$:

$$\mathbf{P}(X_1|X_2, X_3, X_4)\mathbf{P}(X_2|X_3, X_4)\mathbf{P}(X_3|X_4)\mathbf{P}(X_4)$$
$$\mathbf{P}(X_4|X_3, X_2, X_1)\mathbf{P}(X_3|X_2, X_1)\mathbf{P}(X_2|X_1)\mathbf{P}(X_1)$$

If one knows that:

- X_1 is conditionally independent on X_2 and X_4 given X_3: $\mathbf{P}(X_1|X_2, X_3, X_4) = \mathbf{P}(X_1|X_3)$,
- X_2 is conditionally independent on X_3 given X_4: $\mathbf{P}(X_2|X_3, X_4) = \mathbf{P}(X_2|X_4)$,

the **first** expression above can be simplified as:

$$\mathbf{P}(X_1|X_3)\mathbf{P}(X_2|X_4)\mathbf{P}(X_3|X_4)\mathbf{P}(X_4) \ ,$$

whereas the second one **cannot**.

Constructing Bayesian networks

To identify the "best" order among variables before applying the chain rule, two facts can be exploited:

- **causal** knowledge in the form $P(\mathit{effects}|\mathit{cause})$ is often easier to obtain than **diagnostic** knowledge, i.e., $P(\mathit{cause}|\mathit{effects})$
- conditional independence relations turn out to be more easily identifiable in conditional distributions representing **causal** knowledge, i.e., $P(\mathit{effects}|\mathit{cause})$

Constructing Bayesian networks

Example: we have seen that *Toothache* and *Catch* can be considered conditionally independent given *Cavity*:

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

If the chain rule is applied to their full joint distribution as:

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

it can be simplified thanks to the above assumption into:

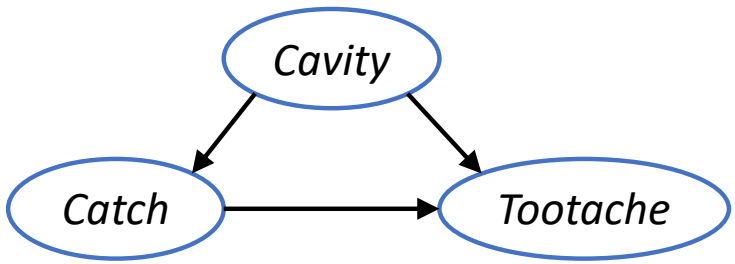
$$\mathbf{P}(\mathit{Toothache}|\mathit{Cavity})\mathbf{P}(\mathit{Catch}|\mathit{Cavity})\mathbf{P}(\mathit{Cavity})$$

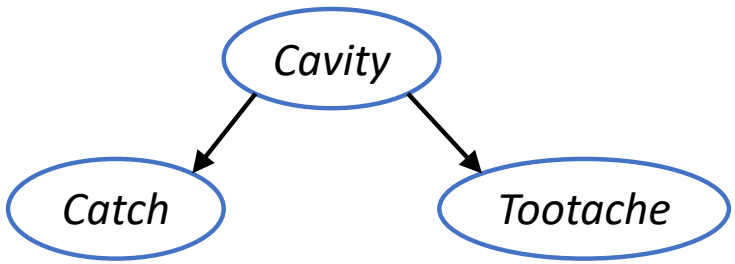
Note that the order of the variables used to apply the chain rule leads to the term in the form $P(\mathit{effects}|\mathit{cause})$ highlighted in red, which allows exploiting the conditional independence assumption.

(cont.)

Constructing Bayesian networks

BNs corresponding to the first, general expression of the full joint distribution $\mathbf{P}(Toothache, Catch, Cavity)$ through the chain rule (left) and to the simplified expression obtained from the conditional independence assumption (right).





Constructing Bayesian networks

The above is an example of the **general rule** to choose the "best" order between the variables when applying the chain rule:

> *select the "root cause" variables first, then the ones they* ***directly*** *influence, and so on, until reaching the variables which have* ***no direct causal influence*** *on the others.*

The same rule allows to **directly** construct a BN by adding nodes one at a time, together with the corresponding edges, without writing the expression of the chain rule first.

Constructing Bayesian networks: an example

You have a new burglar alarm installed at home.

It is fairly reliable at detecting a burglary, but also responds on occasion to minor earthquakes.

You also have two neighbors, John and Mary, who have promised to call you at work when they hear the alarm.

John always calls when he hears the alarm, but sometimes confuses the telephone ringing with the alarm and calls then, too. Mary, on the other hand, likes rather loud music and sometimes misses the alarm altogether.

You may be interested in different probabilistic inferences, e.g.,

estimating the probability of a burglary given the evidence of who has or has not called.

taken from Russell and Norvig, *Artificial Intelligence – A modern Approach*, 2nd Ed., Pearson, 2003

Constructing Bayesian networks: an example

First step: what random variables should be used to represent the events of interest?

A possible choice is the following set of Boolean variables, related to events occurring over any whole day

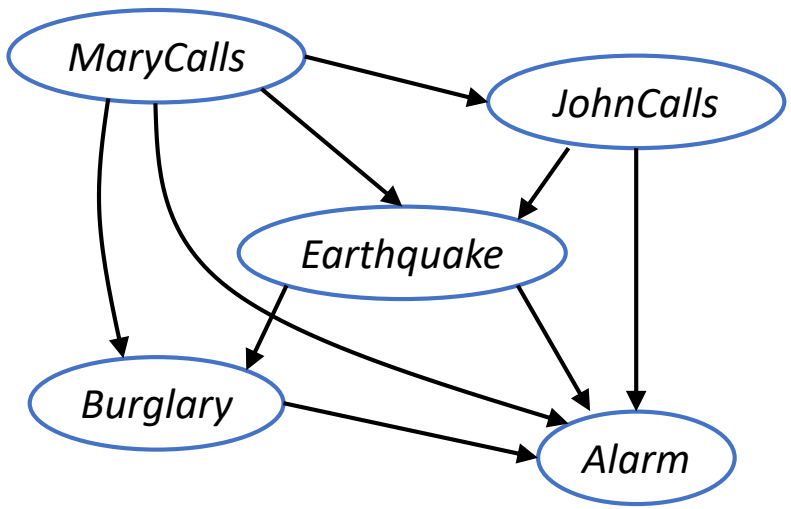
- *Alarm* (A for brevity): whether the alarm sounded or not
- *Burglary* (B): whether a burglary occurred or not
- *Earthquake* (E): whether an earthquake occurred or not
- *JohnCalls* (J): whether John called or not
- *MaryCalls* (M): whether Mary called or not

Constructing Bayesian networks: an example

One of the possible expressions of the full joint distribution obtained using the chain rule is, e.g.:

$$\begin{aligned}&\mathbf{P}(A, B, E, J, M)\\ &\quad = \mathbf{P}(A|B, E, J, M)\mathbf{P}(B|E, J, M)\mathbf{P}(E|J, M)\mathbf{P}(J|M)\mathbf{P}(M)\end{aligned}$$

The corresponding BN, which of course is fully connected:



Constructing Bayesian networks: an example

Let us now try to identify the **causal** dependencies among the five variables, that may lead to **conditional independence assumptions** useful to simplify the BN.

To this aim, one can directly build the BN, adding nodes one at a time, with the coresponding edges, without first writing the corresponding expression of the chain rule.

Burglaries and earthquakes can be considered as "root causes". It can also be assumed that there are no causal dependencies among them. Note that this is only a (reasonable) **assumption**, useful to **simplify** the probabilistic model.

Constructing Bayesian networks: an example

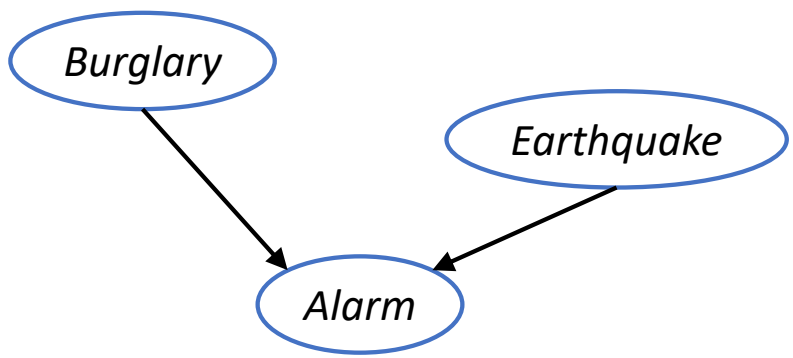
Since there is no causal dependency between burglaries and earthquakes, any of the corresponding node can be added first to the BN, e.g., *Burglary*:

The next node to add is *Earthquake*, with **no** incoming edge from *Burglary*:

Constructing Bayesian networks: an example

Both burglaries and earthquakes **directly** influence the state of the alarm, but **not** the fact that Mary or John calls: we can assume they will call only if they hear (or believe to hear) the alarm sounding.

The next node to add is therefore *Alarm*, with incoming edges from both *Burglary* and *Earthquake*:

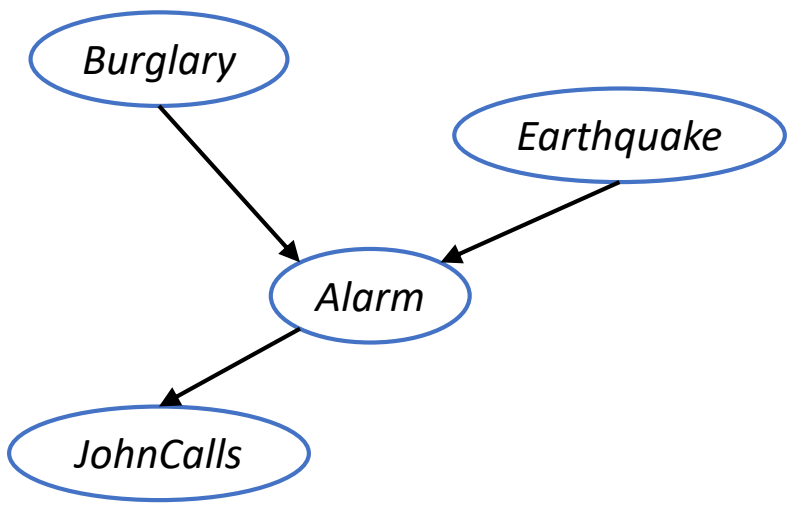


Constructing Bayesian networks: an example

What about *JohnCalls* and *MaryCalls*?

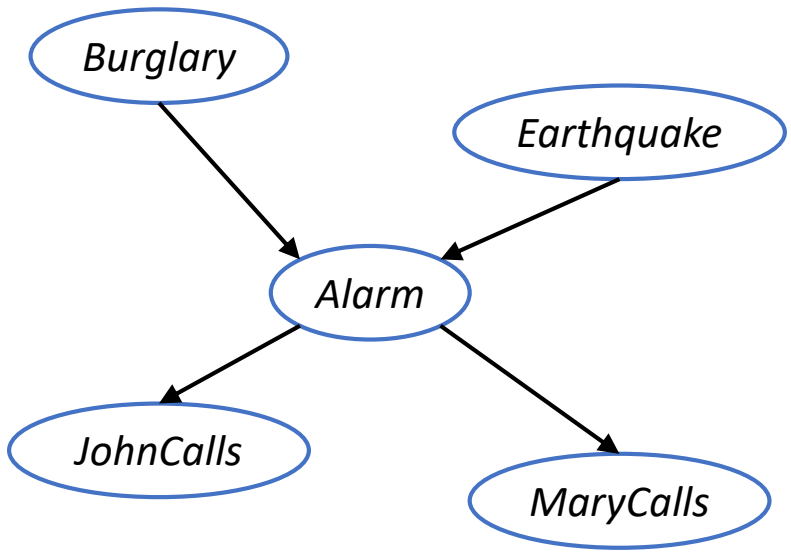
- both variables are **directly** influenced by the state of the alarm, not by burglaries and earthquakes
- we can assume John and Mary do not communicate with each other, so neither variable **directly** influences the other

We can therefore add any of these variables, e.g., *JohnCalls*, with an incoming edge only from *Alarm*:



Constructing Bayesian networks: an example

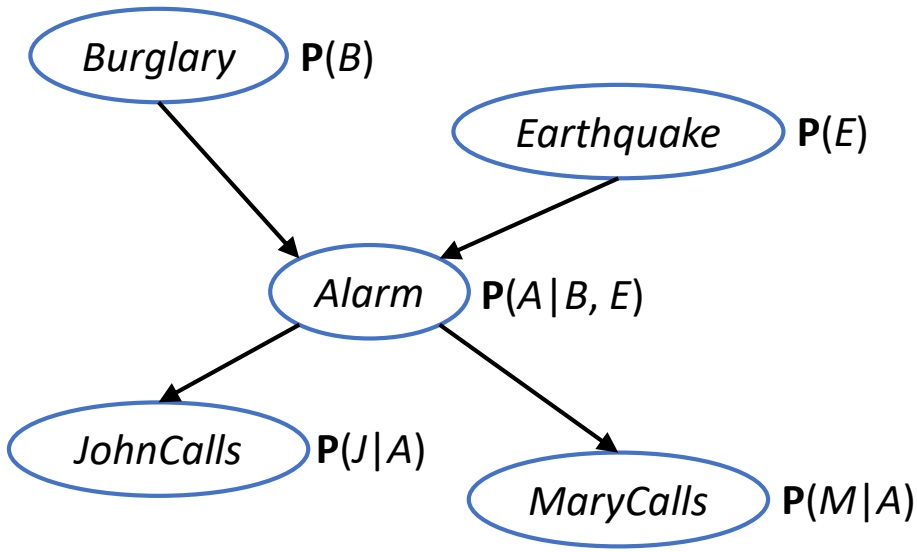
We finally add *MaryCalls*, with an incoming edge only from *Alarm*:



Note that the resulting BN is **not** fully connected as the first one shown above: it contains only 4 edges instead of 10.

Constructing Bayesian networks: an example

It is easy to identify the marginal and conditional distributions associated with each node:



The corresponding expression of the full joint distribution is:

$$\mathbf{P}(B, E, A, J, M) = \mathbf{P}(M|A)\mathbf{P}(J|A)\mathbf{P}(A|E, B)\mathbf{P}(E)\mathbf{P}(B)$$

Constructing Bayesian networks: an example

Is the above BN an **exact** and **complete** probabilistic model of the cosidered domain?

It is likely to be only an **approximate** rather than exact model, due to the **assumptions** made. If the assumptions are reasonable, it might be a good approximation.

What about completeness, e.g., other possible causes of the considered events not represented by the BN? For instance

- no node for Mary's currently listening to loud music
- no node for the telephone ringing and confusing John

Actually all such factors are **implicitly summarised** in the uncertainty associated with the links included in the BN.

Constructing Bayesian networks: an example

The choice of not representing **all** possible factors **explicitly** is an example of **laziness** and **ignorance**: it would be very difficult to consider them all and evaluate their likelihood.

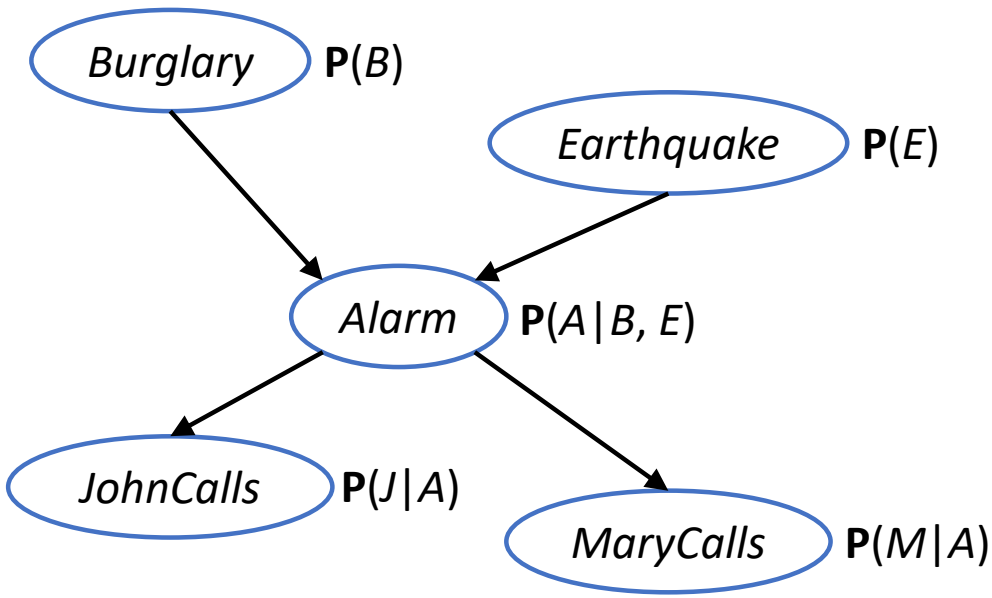
On the other hand, the distributions associated to a BN **summarise** a **potentially infinite** set of circumstances in which the considered events can happen or not, e.g.:

- $\mathbf{P}(A|E,B)$ summarises all the other causes, beside earthquakes and burglaries, that make the alarm sound or fail to sound (passing helicopter, high humidity, power failure, dead battery, cut wires, a dead mouse stuck inside the bell, etc.)
- $\mathbf{P}(J|A)$ and $\mathbf{P}(M|A)$ summarises all the other causes, beside the state of the alarm, that make John or Mary call or fail to call (out to lunch, on vacation, listening to loud music, passing helicopter, etc.).

This is the way in which probabilistic models can be kept small enough to be **practically useful**, albeit only **approximate**.

Constructing Bayesian networks: an example

What are the conditional independence assumptions **informally** made when constructing our BN?



Constructing Bayesian networks: an example

To identify conditional independence assumptions, consider

- the **general** expression of the full joint distribution using the chain rule for the chosen order between the variables:

$$\begin{aligned}&\mathbf{P}(B, E, A, J, M)\\ &\quad = \mathbf{P}(M|J, A, E, B)\mathbf{P}(J|A, E, B)\mathbf{P}(A|E, B)\mathbf{P}(E|B)\mathbf{P}(B)\end{aligned}$$

- the expression corresponding to the BN:

$$\mathbf{P}(B, E, A, J, M) = \mathbf{P}(M|A)\mathbf{P}(J|A)\mathbf{P}(A|E, B)\mathbf{P}(E)\mathbf{P}(B)$$

(cont.)

Constructing Bayesian networks: an example

Comparing the corresponding terms in the two expressions above:

1. *Earthquake* is (unconditionally) independent on *Burglary*:

$$\mathbf{P}(E|B) = \mathbf{P}(E)$$

2. *JohnCalls* is conditionally independent on *Burglary* and *Earthquake* given *Alarm*:

$$\mathbf{P}(J|A, E, B) = \mathbf{P}(J|A)$$

3. *MaryCalls* is conditionally independent on *Burglary*, *Earthquake* and *JohnCalls* given *Alarm*:

$$\mathbf{P}(M|J, A, E, B) = \mathbf{P}(M|A)$$

Constructing Bayesian networks: an example

The ones listed above are the **only** conditional (or absolute) independence assumptions **explicitly** made during the construction of the considered BN.

Nevertheless, it can be shown that a BN also implies other conditional independence assumptions (see later).

Attention should be paid to avoid drawing **wrong** conclusions about independence relations implied by a given BN

- for instance, from the fact that there are no incoming links to *Burglary*, it would be **not correct** to conclude that *Burglary* is independent on all the other variables

Constructing Bayesian networks: an example

To define the full joint distribution $\mathbf{P}(B, E, A, J, M)$, $2^5 - 1 = 31$ probability values must be specified.

How much do the above conditional independence assumptions simplify its definition?

- $\mathbf{P}(M|A)$: $2 \times (2^1 - 1) = 2$ values,
- $\mathbf{P}(J|A)$: $2 \times (2^1 - 1) = 2$ values,
- $\mathbf{P}(A|E, B)$: $2^2 \times (2^1 - 1) = 4$ values,
- $\mathbf{P}(E)$: $2^1 - 1 = 1$ value,
- $\mathbf{P}(B)$: $2^1 - 1 = 1$ value,

for a total of **10** probability values, instead of **31**.

Constructing Bayesian networks: an example

To construct the above BN the variables have been ordered from "root causes" to "end effects": *Burglary*, *Earthquake*, *Alarm*, *JohnCalls*, *MaryCalls*.

As a consequence, **all** the distributions associated to the nodes of the BN (i.e., to the corresponding expression of the chain rule) are of **causal** form, $P(\mathit{effect}|\mathit{causes})$, which are usually simpler to estimate, and allow conditional independence relations to be exploited.

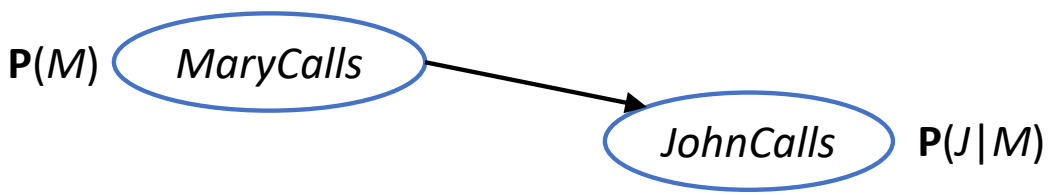
What happens if one choses a different order among the variables to construct a BN?

Constructing Bayesian networks: an example

Consider for instance adding variables in the following order: *MaryCalls*, *JohnCalls*, *Alarm*, *Burglary*, *Earthquake*.

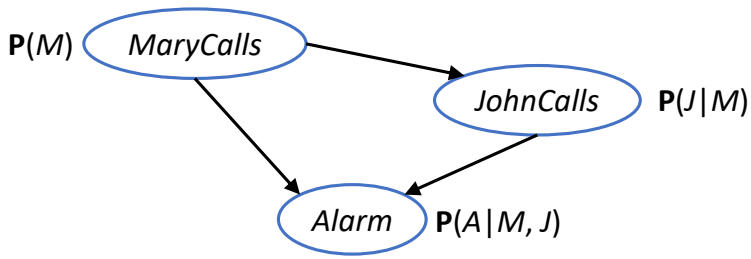
- Adding *MaryCalls*:

- Adding *JohnCalls*: if Mary calls, it is likely the alarm has sounded, which makes it more likely that John calls: *JohnCalls* needs *MaryCalls* as a parent

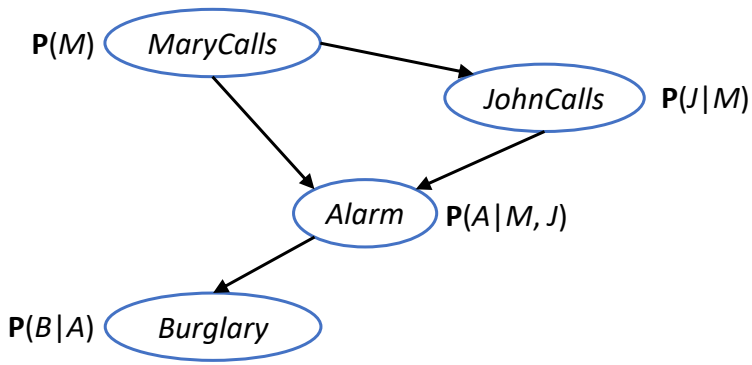


Constructing Bayesian networks: an example

- Adding *Alarm*: if both call, it is more likely that the alarm has sounded than if just one or neither call: both *MaryCalls* and *JohnCalls* are needed as parents

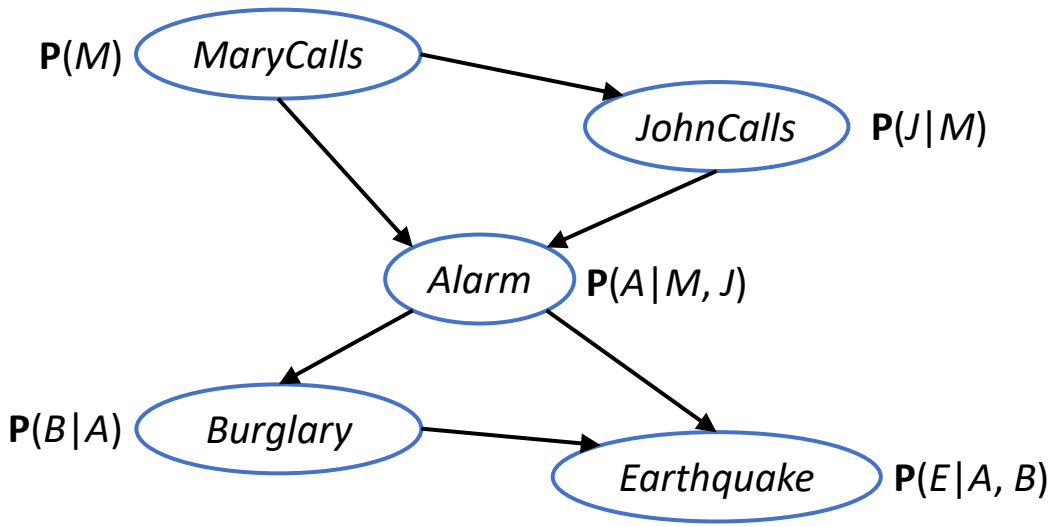


- Adding *Burglary*: if we know the alarm is sounding, a call from John or Mary does not provide additional information about burglary: only *Alarm* is needed as parent



Constructing Bayesian networks: an example

- Adding *Earthquake*: if the alarm is sounding, it is more likely that an earthquake occurred; if we know that also a burglary occurred, this explains the alarm, and the probability of an earthquake would be only slightly above normal; a call from John or Mary does not provide any additional information on earthquakes: only *Alarm* and *Burglary* are needed as parents



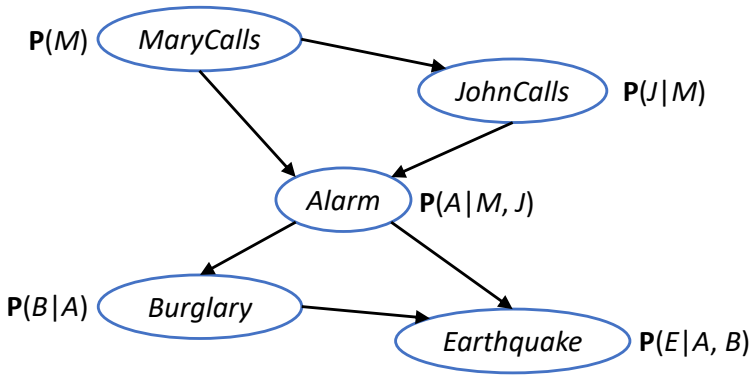
Constructing Bayesian networks: an example

The resulting BN has two more edges than the previous one, and therefore requires **more** probability values to be specified.

Moreover, the required distributions are much difficult to estimate, since they require unnatural probability judgments, and in particular **non-causal** ones, e.g., assessing the probability of an earthquake given that a burglary occurred and the alarm sounded, $\mathbf{P}(E|B, A)$.

Exercises

1. Write the expression of the full joint distribution using the chain rule, corresponding to the BN on the right



2. Identify the conditional independence assumptions implicitly made during the construction of the above BN
3. Compute the number of probability values to be specified to define the distributions associated to the above BN
4. Construct a BN using a different order among the variables: *MaryCalls*, *JohnCalls*, *Earthquake*, *Burglary*, *Alarm*, and repeat steps 1–3