

# Introduction to Bayesian Networks

Diogo Pernes

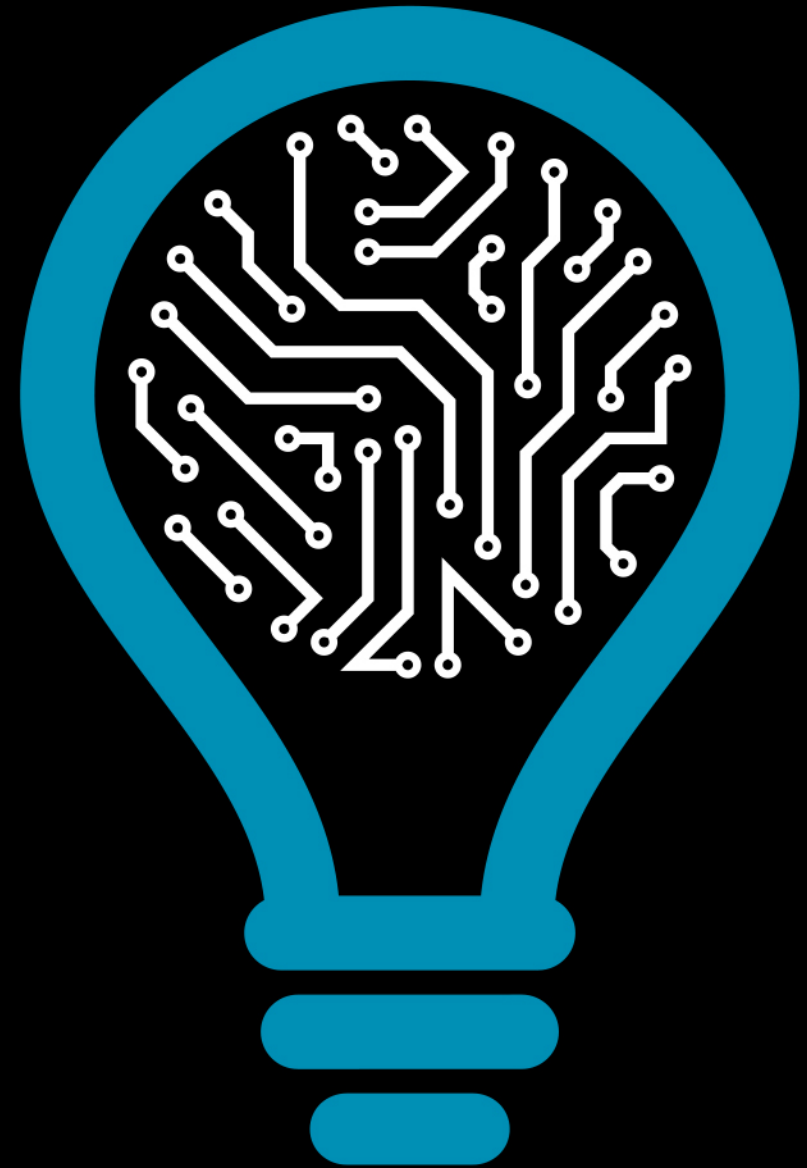
INESC TEC & University of Porto

[diogo.pernes@inesctec.pt](mailto:diogo.pernes@inesctec.pt)

VISUM Summer School 2021



INSTITUTE FOR SYSTEMS  
AND COMPUTER ENGINEERING,  
TECHNOLOGY AND SCIENCE



# Probability: a brief review

# Probability

Def.: Given a (discrete) random variable  $X \in \mathcal{X} = \{x^{(1)}, x^{(2)}, \dots, x^{(d)}\}$ , a function  $P(X) : \mathcal{X} \mapsto [0, 1]$  is a *probability distribution* for  $X$  if it satisfies:

$$\sum_{i=1}^d P(X = x^{(i)}) = 1 \quad \text{or, in shorthand notation,} \quad \sum_X P(X) = 1.$$

Def.: Given two (discrete) random variables  $X \in \mathcal{X}$  and  $Y \in \mathcal{Y}$  with probability distributions  $P(X)$  and  $P(Y)$ , respectively, the function  $P(X, Y) : \mathcal{X} \times \mathcal{Y} \mapsto [0, 1]$  is a *joint probability distribution* for  $X$  and  $Y$  if it satisfies:

$$\sum_X P(X, Y) = P(Y) \quad \text{and} \quad \sum_Y P(X, Y) = P(X).$$

# Conditional probability, independence and Bayes' theorem

Def.: Given two random variables  $X$  and  $Y$ , the *conditional probability* of  $Y$  given  $X$  is:

$$P(Y \mid X) = \frac{P(X, Y)}{P(X)}.$$

Thm. (Bayes): Given two random variables  $X$  and  $Y$ , the following equality holds:

$$P(Y \mid X)P(X) = P(X \mid Y)P(Y).$$

Def.:  $X$  and  $Y$  are *independent* if  $P(X, Y) = P(X)P(Y)$ . Notation:  $(X \perp Y)$

Thm.: The following equalities are equivalent:

$$P(X, Y) = P(X)P(Y), \quad P(Y \mid X) = P(Y), \quad P(X \mid Y) = P(X).$$

# Conditional independence

Def.:  $X$  and  $Y$  are *conditionally independent* given  $Z$  if  $P(X, Y \mid Z) = P(X \mid Z)P(Y \mid Z)$

Notation:  $(X \perp Y \mid Z)$

Thm.: The following equalities are equivalent:

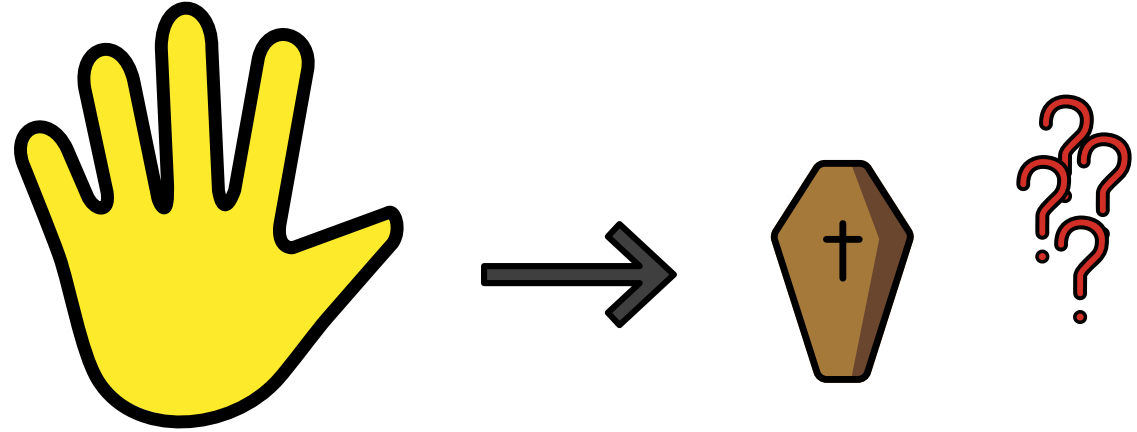
$$P(X, Y \mid Z) = P(X \mid Z)P(Y \mid Z), \quad P(Y \mid X, Z) = P(Y \mid Z), \quad P(X \mid Y, Z) = P(X \mid Z).$$

Remark: Marginal independence and conditional independence do *not* imply each other!

$$(X \perp Y \mid Z) \not\Rightarrow (X \perp Y) \\ (X \perp Y) \not\Rightarrow (X \perp Y \mid Z)$$

# Marginal independence vs. Conditional independence

$$(X \perp Y \mid Z) \not\Rightarrow (X \perp Y)$$



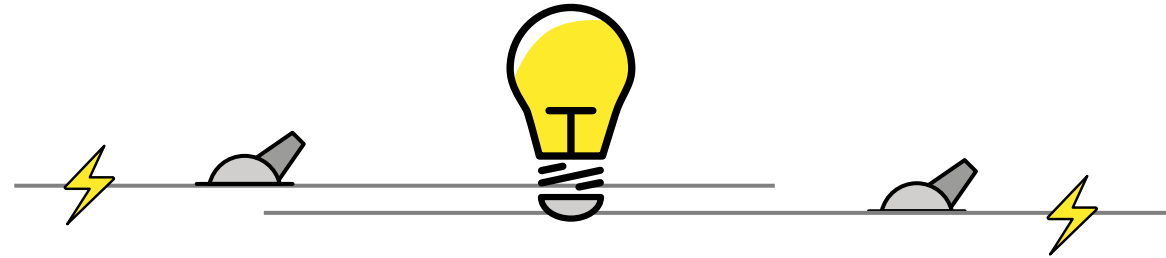
Palm size ( $X$ ) and life expectancy ( $Y$ ) are statistically associated (i.e. not independent): the larger your hand is, the less years you are expected to live!

The association disappears if you group the data by gender ( $Z$ ): women tend to have smaller hands and to live longer than men.

Thus, palm size and life expectancy are not marginally independent, but they are conditionally independent given the gender.

# Marginal independence vs. Conditional independence

$$(X \perp Y) \not\Rightarrow (X \perp Y \mid Z)$$



You have two switches  $X$  and  $Y$  controlling the same light bulb ( $Z$ ), which are switched independently from one another ( $X \perp Y$ ). The light is off if and only if both switches are in the off position.

If you observe that the light is on and one of the switches is in the off position, then you immediately conclude that the other switch must be in the on position.

Thus, the switches states are marginally independent but they are not conditionally independent given the light state.

# Table conditional probability distributions

- Throughout this presentation, we shall in general consider random variables that are discrete and have finite cardinality.
  - Some loss of generality.
  - Most concepts may be generalized to any random variable with no effort.
  - Facilitates exposition and understanding (hopefully).
- A general and convenient way to represent and parameterize a discrete conditional probability distribution (CPD) where all variables have finite cardinality is through a table CPD.

E.g.:

$$X \in \{x^{(1)}, x^{(2)}\}$$

$$Y \in \{y^{(1)}, y^{(2)}, y^{(3)}\}$$

$P(Y \mid X)$	$x^{(1)}$	$x^{(2)}$
$y^{(1)}$	0.6	0.8
$y^{(2)}$	0.1	0.2
$y^{(3)}$	0.3	0



# Joint distributions and Bayesian networks

# From factorizations to DAGs

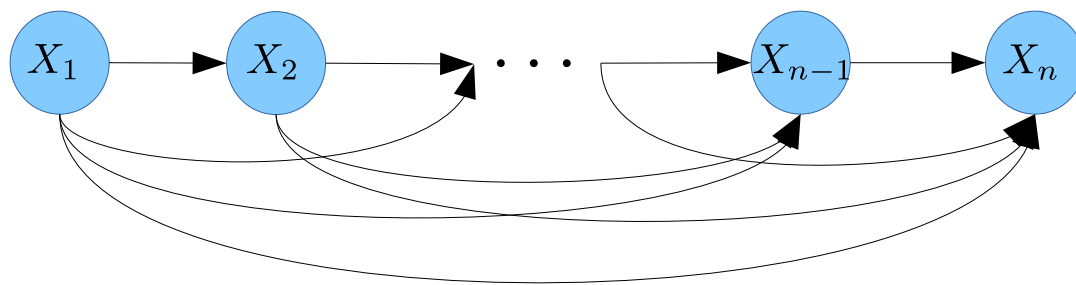
$$\begin{aligned} P(X_1, X_2, \dots, X_n) &= P(X_1)P(X_2, X_3, \dots, X_n \mid X_1) \\ &= P(X_1)P(X_2 \mid X_1)P(X_3, X_4, \dots, X_n \mid X_1, X_2) \\ &= \dots \\ &= \underbrace{P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_1, X_2) \cdots P(X_n \mid X_1, X_2, \dots, X_{n-1})}_{\text{chain rule of probability}} \end{aligned}$$

# From factorizations to DAGs

$$P(X_1, X_2, \dots, X_n) = P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_1, X_2) \cdots P(X_n \mid X_1, X_2, \dots, X_{n-1})$$

Given the factorization above, let us build a *directed graph*  $\mathcal{G}$  such that:

- There is one vertex for each  $X_i$ ;
- There is one edge  $X_i \rightarrow X_j$  if there exists a factor where  $X_j$  is conditioned on  $X_i$ .



$\mathcal{G}$  is a *complete directed acyclic graph* (DAG)

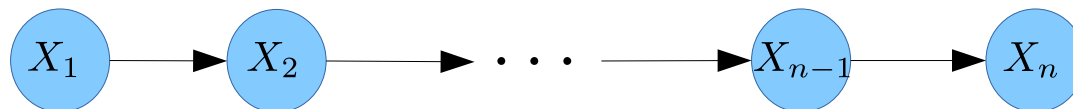


## From factorizations to DAGs

Now, let us assume that  $(X_i \perp X_1, X_2, \dots, X_{i-2} \mid X_{i-1})$ ,

hence  $P(X_i \mid X_1, X_2, \dots, X_{i-1}) = P(X_i \mid X_{i-1})$  and the former factorization and DAG reduce to:

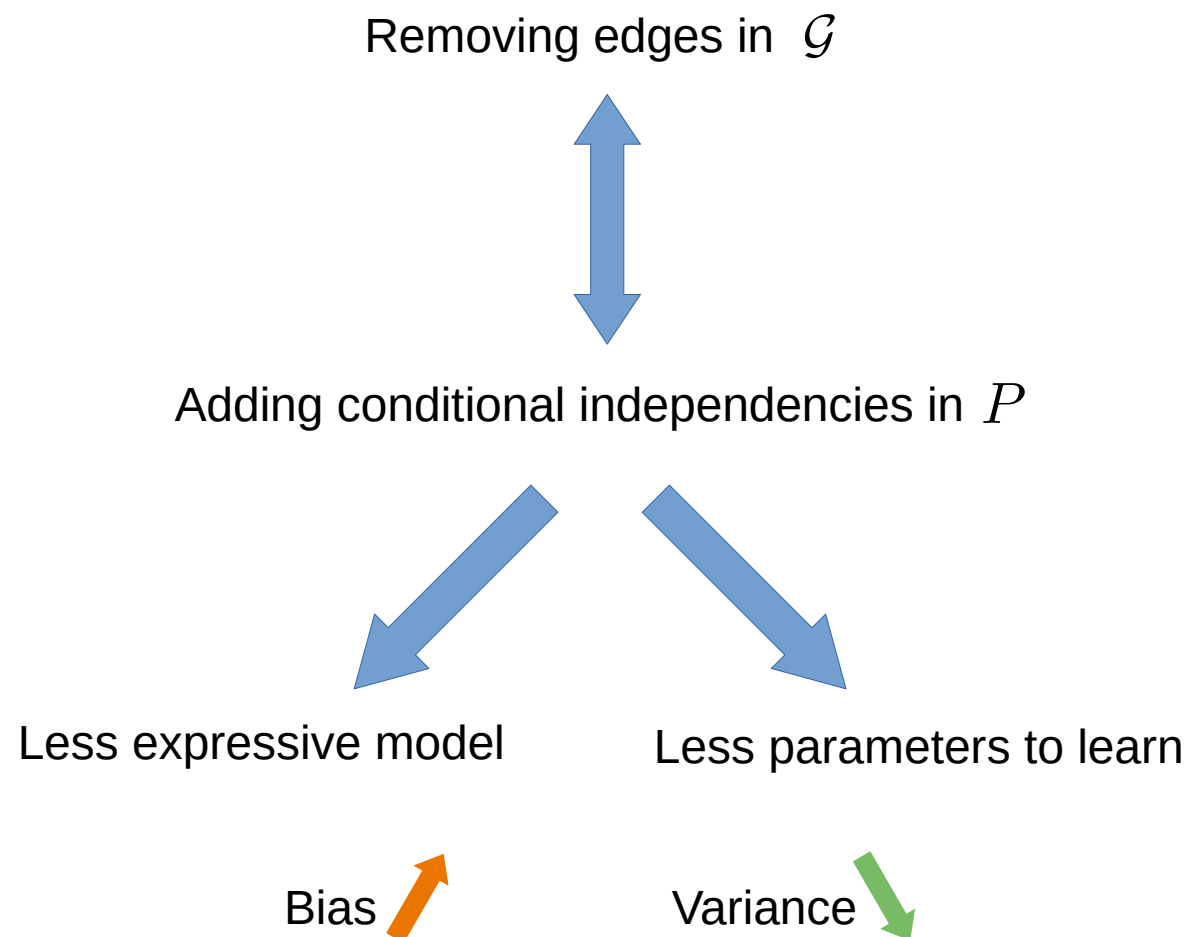
$$P(X_1, X_2, \dots, X_n) = P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_2) \cdots P(X_n \mid X_{n-1})$$



$\mathcal{G}$  is now a sparser DAG

A DAG corresponding to a factorization of a joint probability distribution is known as a *Bayesian network*.

# Sparser vs. denser DAGs



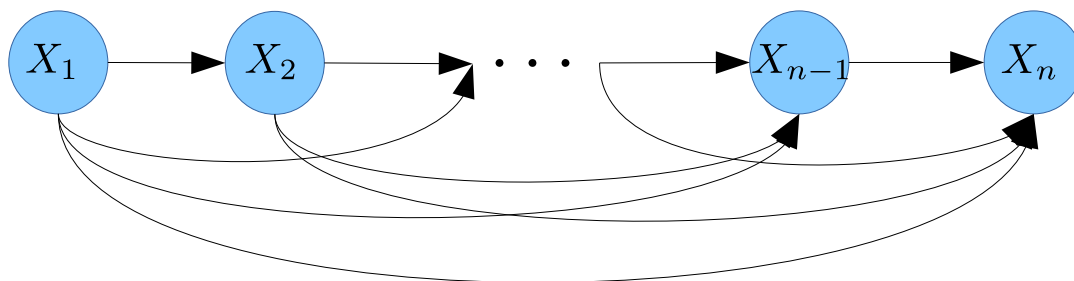
# Sparser vs. denser DAGs

Why sparser means less expressive?

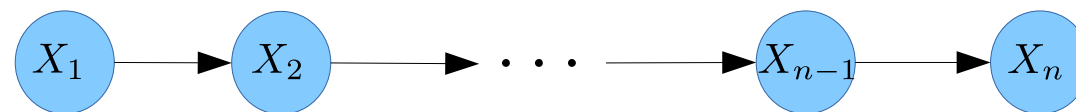
- The model can't capture dependencies between variables that it assumes to be conditionally independent.

Why sparser means with less parameters?

- Removing an edge corresponds to removing a column in the table CPD of the corresponding factor.
- e.g.: Assume each  $X_i$  takes  $d$  possible values and consider the two examples below.



Complete DAG:  $d^n - 1$  parameters



Chain:  $O(d^2 n)$  parameters

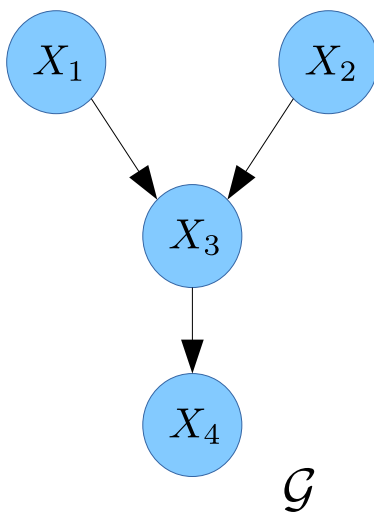
# From DAGs to factorizations

Given a DAG  $\mathcal{G}$ , we find the corresponding factorization of the joint distribution by applying the following formula:

$$P(X_1, X_2, \dots, X_n) = \prod_{i=1}^n P(X_i \mid \text{Pa}_{\mathcal{G}}(X_i)),$$

where  $\text{Pa}_{\mathcal{G}}(X_i)$  are the parent nodes of  $X_i$  in  $\mathcal{G}$ .

e.g.:



$$P(X_1, X_2, X_3, X_4) = P(X_1)P(X_2)P(X_3 \mid X_1, X_2)P(X_4 \mid X_3)$$

Wait... But why do we care  
about DAGs at all?





## Mostly, for two reasons...

1. Because DAGs are great tools for probabilistic and causal reasoning. In the context of machine learning, they provide an elegant framework to incorporate prior knowledge into the model.

Solving a probabilistic problem be like...



*before drawing the DAG*



*after drawing the DAG*

2. Because of d-separation: a simple graphical criterion to find the conditional independencies implied by a given factorization of a joint distribution.

# Probabilistic reasoning, an example: the two child problem

- Mrs. Smith has two children.
- You ask her if she has a son.
- She answers 'yes'.

Q: What is the probability that both children are boys?

Spoiler: It is *not*  $1/2$ !













# Probabilistic reasoning, an example: the two child problem

- Mrs. Smith has two children.
- You ask her if she has a son.
- She answers 'yes'.

Q: What is the probability that both children are boys?

A:  $1/3$

Older child	Younger child	Mrs. Smith's answer
		No
		Yes
		Yes
		Yes

We got the right answer by enumerating all possible cases, but is it intuitive?

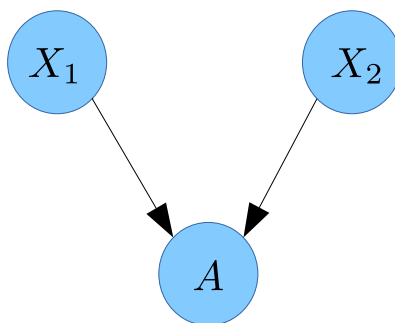
# Probabilistic reasoning, an example: the two child problem

Let us draw a DAG with three random variables:

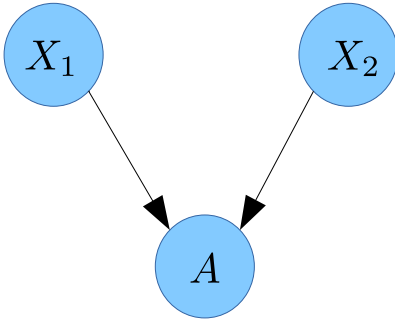
- $X_1 \in \{\text{girl}, \text{boy}\}$ , representing the sex of the older child;
- $X_2 \in \{\text{girl}, \text{boy}\}$ , representing the sex of the younger child;
- $A \in \{\text{yes}, \text{no}\}$ , representing Mrs. Smith's answer to the question.

Now, we draw edges representing how the three random variables influence each other:

- $A$  depends on both  $X_1$  and  $X_2$ :  $A = \text{no}$  (with prob. 1) if both children are girls, and  $A = \text{yes}$  otherwise (also with prob. 1). Thus, we draw edges  $X_1 \rightarrow A$  and  $X_2 \rightarrow A$ .
- The sex of a child does not influence the sex of the other, so  $X_1$  and  $X_2$  are not directly connected.



# Probabilistic reasoning, an example: the two child problem



$$P(X_1, X_2, A) = P(X_1)P(X_2)P(A \mid X_1, X_2)$$

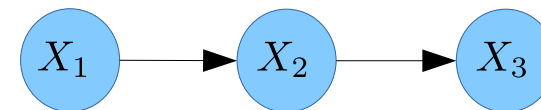
$$\begin{aligned} P(X_1 = \text{boy}, X_2 = \text{boy} \mid A = \text{yes}) &= \frac{P(X_1 = \text{boy}, X_2 = \text{boy}, A = \text{yes})}{P(A = \text{yes})} \\ &= \frac{P(X_1 = \text{boy})P(X_2 = \text{boy})P(A = \text{yes} \mid X_1 = \text{boy}, X_2 = \text{boy})}{\sum_{X_1, X_2} P(X_1, X_2, A = \text{yes})} \\ &= \frac{1/2 \times 1/2 \times 1}{1/2 \times 1/2 \times 1 + 1/2 \times 1/2 \times 1 + 1/2 \times 1/2 \times 1} = \frac{1/4}{3/4} = \frac{1}{3} \end{aligned}$$

# Chains and forks

Given 3 random variables  $X_1, X_2$  and  $X_3$ , assume that  $(X_1 \perp X_3 \mid X_2)$ .

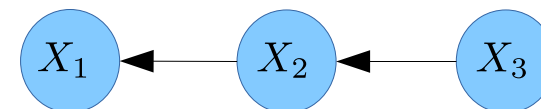
Then,

$$P(X_1, X_2, X_3) = P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_2)$$



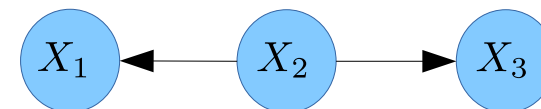
a chain

$$= P(X_3)P(X_2 \mid X_3)P(X_1 \mid X_2)$$



a reversed chain

$$= P(X_1 \mid X_2)P(X_2)P(X_3 \mid X_2)$$



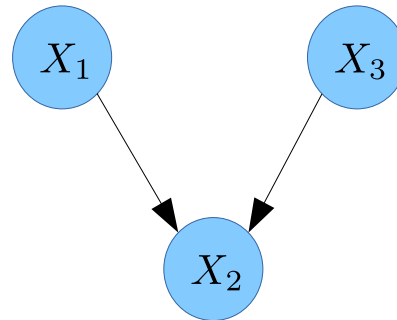
a fork

# Immoralities

Given 3 random variables  $X_1, X_2$  and  $X_3$ , assume that  $(X_1 \perp X_3)$ .

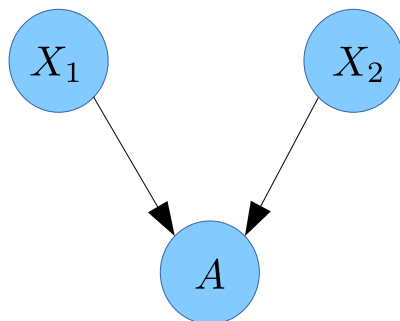
Then,

$$P(X_1, X_2, X_3) = P(X_1)P(X_3)P(X_2 \mid X_1, X_3)$$



an immorality, vertex  
 $X_2$  is the *collider*

## The two child problem revisited



$$(X_1 \perp X_2), \quad (X_1 \not\perp X_2 \mid A)$$









- Before we observe Mrs. Smith answer, the sex of the younger child is independent of the sex of the older child.
- When we observe Mrs. Smith answer, we immediately find that it is impossible that both children are girls, so the sexes of the two children become dependent from each other.
- This association is *non-causal*, and this is why most people find it non-intuitive.



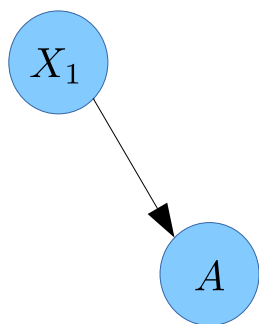
# The two child problem revisited

- Mrs. Smith has two children.
- You ask her if her *older child* is a boy.
- She answers 'yes'.

Q: What is the probability that both children are boys?

Older child	Younger child	Mrs. Smith's answer
		No
		No
		Yes
		Yes

A: 1/2



$$\begin{aligned}
 P(X_1 = \text{boy}, X_2 = \text{boy} \mid A = \text{yes}) &= \frac{P(X_1 = \text{boy}, X_2 = \text{boy}, A = \text{yes})}{P(A = \text{yes})} \\
 &= \frac{1/4}{2/4} = \frac{1}{2}
 \end{aligned}$$

$$(X_1 \perp X_2), \quad (X_1 \perp X_2 \mid A)$$

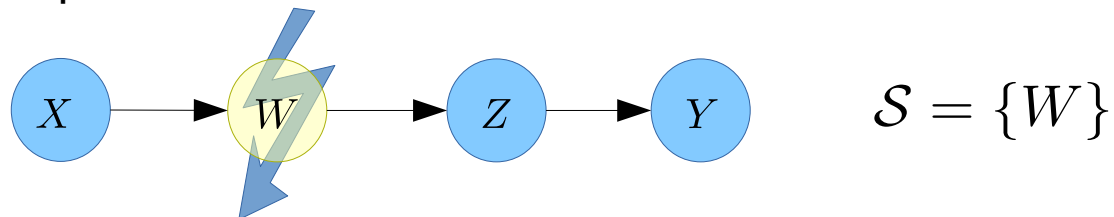
$$P(X_1, X_2, A) = P(X_1)P(X_2)P(A \mid X_1)$$

# Cond. independencies in larger graphs: d-separation

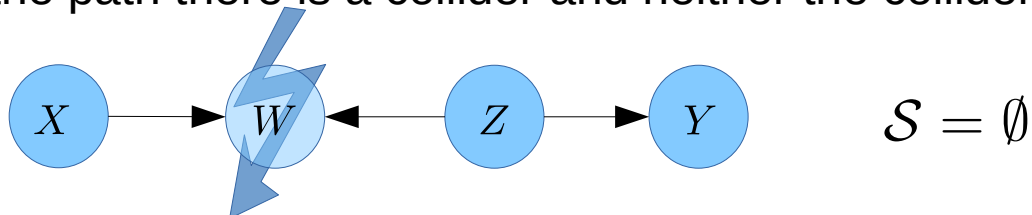
*Blocked paths:*

An undirected path between two vertices  $X$  and  $Y$  is *blocked* by a (potentially empty) conditioning set  $\mathcal{S}$  if at least one of the following conditions holds:

- Along the path there is a chain or a fork which includes at least one node in  $\mathcal{S}$ .



- Along the path there is a collider and neither the collider nor any of its descendants are in  $\mathcal{S}$ .



# Cond. independencies in larger graphs: d-separation

*d-separation:*

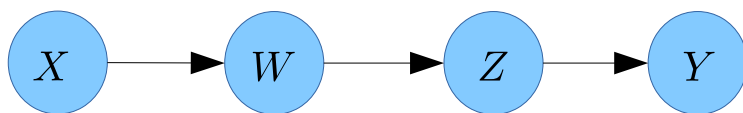
Two nodes  $X$  and  $Y$  are *d-separated* by a (potentially empty) conditioning set  $\mathcal{S}$  if conditioning on  $\mathcal{S}$  blocks all paths between  $X$  and  $Y$ .

Remarkably,

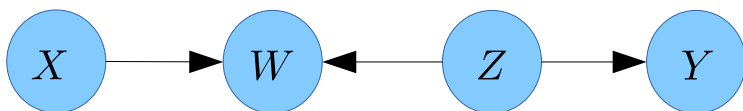
$$X \text{ and } Y \text{ d-separated by } \mathcal{S} \longrightarrow (X \perp Y \mid \mathcal{S})$$

# Cond. independencies in larger graphs: d-separation

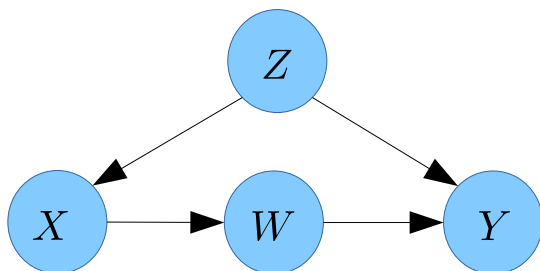
A few examples...



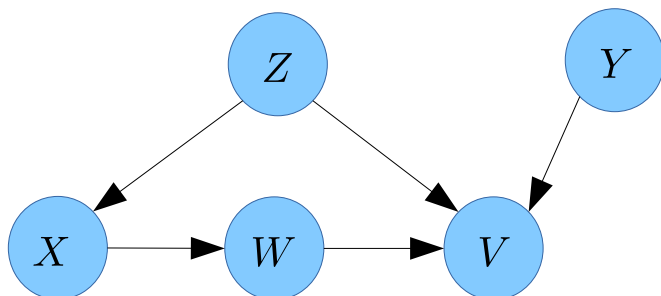
$$(X \perp Y \mid W), \quad (X \perp Y \mid Z)$$



$$(X \perp Y), \quad (X \perp Y \mid Z), \quad (X \perp Y \mid W, Z)$$



$$(X \perp Y \mid W, Z)$$



$$(X \perp Y), (X \perp Y \mid W), (X \perp Y \mid Z), \\ (X \perp Y \mid W, Z), (X \perp Y \mid W, Z, V)$$

# Learning Bayesian networks



# The learning problem

Key assumption: Our dataset  $\mathcal{D} = \{x[1], \dots, x[m]\}$  consists of  $m$  *independent and identically distributed* samples, drawn from an unknown distribution  $P$ .


Goal: Find a “good” approximation  $\hat{P}$  of the true underlying distribution  $P$ .


How? Two approaches:

- a) Choose a fixed graph structure  $\mathcal{G}$  and learn parameters  $\theta^*$  such that the obtained distribution  $\hat{P} = P_{\mathcal{G}}(\theta^*)$  is “close” to  $P$  (e.g.: learning a naive Bayes model, a hidden Markov model, etc.).
- b) Learn a graph structure  $\mathcal{G}^*$  and parameters  $\theta^*$  such that the obtained distribution  $\hat{P} = P_{\mathcal{G}^*}(\theta^*)$  is “close” to  $P$  (e.g.: causal discovery).

# The learning problem

Usually, our objective is  $L(\mathcal{G}, \theta) = \frac{1}{m} \sum_{i=1}^m \log P_{\mathcal{G}}(x[i] : \theta)$  and the optimization problem is either:

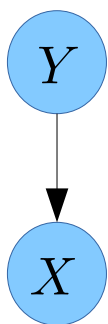
1.  $\max_{\theta} L(\mathcal{G}, \theta)$  for a fixed  $\mathcal{G}$ ,
2. or  $\max_{\mathcal{G}, \theta} L(\mathcal{G}, \theta)$ .  not covered in this presentation

The optimal distribution  $\hat{P}$  can be viewed as maximizing the probability (likelihood) that our dataset  $\mathcal{D}$  was drawn from it.  maximum likelihood estimation (MLE)

# MLE for Bayesian networks: complete data

- In case every random variable is observed in every data sample, finding the ML parameters for a Bayesian network is very easy!
- It can be proven that the optimal parameters correspond to the *empirical probabilities* obtained from the provided dataset.

e.g.:  $X \in \{0, 1\}, \quad Y \in \{0, 1\},$



$P(Y)$		
$Y = 0$	$\theta_0$	
$Y = 1$	$\theta_1$	
$P(X   Y)$	$Y = 0$	$Y = 1$
$X = 0$	$\theta_{0 0}$	$\theta_{0 1}$
$X = 1$	$\theta_{1 0}$	$\theta_{1 1}$

$\mathcal{D} = \{(x[i], y[i])\}_i = \{(0, 1), (0, 0), (1, 1), (1, 0), (1, 1)\}$

$$\theta_0 = \frac{2}{5} = .4 \quad \theta_1 = \frac{3}{5} = .6$$

$$\theta_{0|0} = \frac{1}{2} = .5 \quad \theta_{0|1} = \frac{1}{3} = .333\dots$$

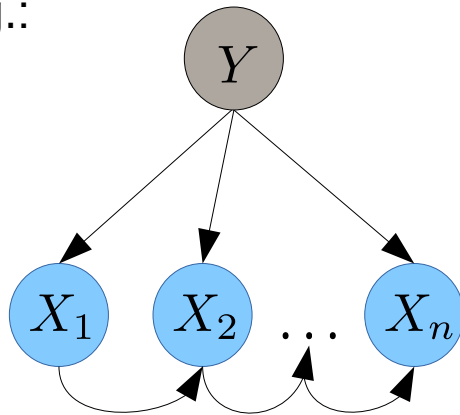
$$\theta_{1|0} = \frac{1}{2} = .5 \quad \theta_{1|1} = \frac{2}{3} = .666\dots$$



# MLE for Bayesian networks: missing data

- In some situations, there are some variables that are not observed in the data – these are called *latent variables*.
- This leads to multiple *local optima* in the objective function.

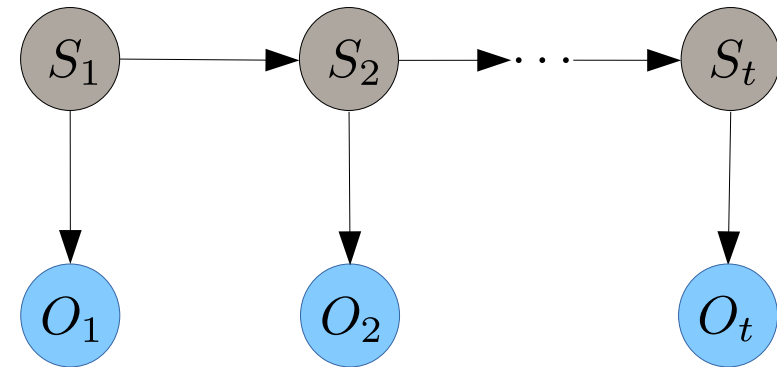
e.g.:



$$\mathcal{D} = \{(x_1[i], \dots, x_n[i])\}_i$$

( $y[i]$  's not observed)

Bayesian clustering




$$\mathcal{D} = \{(o_1[i], \dots, o_t[i])\}_i$$

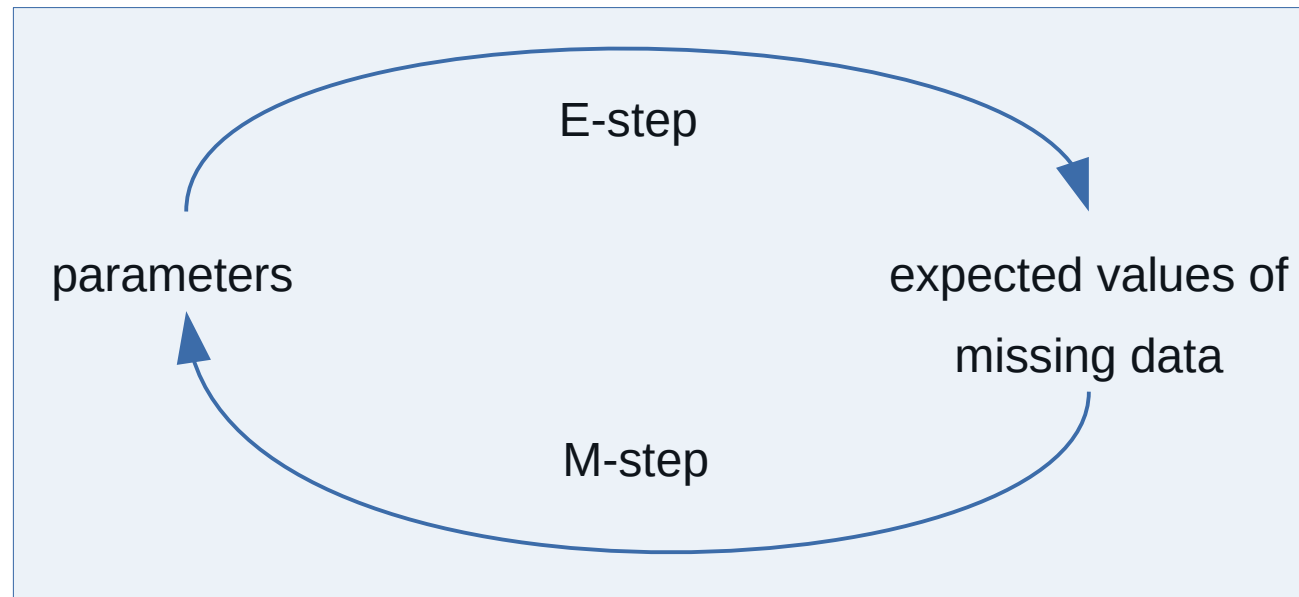
( $s_1[i], \dots, s_t[i]$  's not observed)

Hidden Markov model (HMM)

# MLE for Bayesian networks: missing data

In this situation we have two options (algorithms) to train our model:

1. Gradient ascent over the likelihood function.
2. Expectation Maximization (EM).  faster convergence



## Want to learn more?

- Probabilistic Graphical Models Specialization, Coursera, Stanford University – <https://www.coursera.org/specializations/probabilistic-graphical-models>
- Koller & Friedman, Probabilistic Graphical Models Principles and Techniques, MIT Press, 2009

