

# From Independencies to Directed Graphs

Michael Gutmann

Probabilistic Modelling and Reasoning (INFR11134)  
School of Informatics, University of Edinburgh

Spring Semester 2020

# Recap

- ▶ We talked about reasonably weak assumption to facilitate the efficient representation of a probabilistic model
- ▶ Independence assumptions reduce the number of interacting variables
- ▶ Parametric assumptions restrict the way the variables may interact.
- ▶ (Conditional) independence assumptions lead to a factorisation of the pdf/pmf, e.g.
  - ▶  $p(\mathbf{x}, \mathbf{y}, \mathbf{z}) = p(\mathbf{x})p(\mathbf{y})p(\mathbf{z})$
  - ▶  $p(x_1, \dots, x_d) = p(x_d | x_{d-3}, x_{d-2}, x_{d-1})p(x_1, \dots, x_{d-1})$

# Program

1. Equivalence of factorisation and ordered Markov property
2. Understanding models from their factorisation

# Program

1. Equivalence of factorisation and ordered Markov property
  - Chain rule
  - Ordered Markov property implies factorisation
  - Factorisation implies ordered Markov property
2. Understanding models from their factorisation

# Chain rule

Iteratively applying the product rule allows us to factorise any joint pdf (pmf)  $p(\mathbf{x}) = p(x_1, x_2, \dots, x_d)$  into product of conditional pdfs.

$$\begin{aligned} p(\mathbf{x}) &= p(x_1)p(x_2, \dots, x_d|x_1) \\ &= p(x_1)p(x_2|x_1)p(x_3, \dots, x_d|x_1, x_2) \\ &= p(x_1)p(x_2|x_1)p(x_3|x_1, x_2)p(x_4, \dots, x_d|x_1, x_2, x_3) \\ &\vdots \\ &= p(x_1)p(x_2|x_1)p(x_3|x_1, x_2) \dots p(x_d|x_1, \dots, x_{d-1}) \\ &= p(x_1) \prod_{i=2}^d p(x_i|x_1, \dots, x_{i-1}) \\ &= \prod_{i=1}^d p(x_i|\text{pre}_i) \end{aligned}$$

with  $\text{pre}_i = \text{pre}(x_i) = \{x_1, \dots, x_{i-1}\}$ ,  $\text{pre}_1 = \emptyset$  and  $p(x_1|\emptyset) = p(x_1)$

The chain rule can be applied to any ordering  $x_{k_1}, \dots, x_{k_d}$ . Different orderings give different factorisations.

# From (conditional) independence to factorisation

$p(\mathbf{x}) = \prod_{i=1}^d p(x_i | \text{pre}_i)$  for the ordering  $x_1, \dots, x_d$

- ▶ For each  $x_i$ , we condition on all previous variables in the ordering.
- ▶ Assume that, for each  $i$ , there is a minimal subset of variables  $\pi_i \subseteq \text{pre}_i$  such that  $p(\mathbf{x})$  satisfies

$$x_i \perp\!\!\!\perp (\text{pre}_i \setminus \pi_i) \mid \pi_i$$

for all  $i$ .

- ▶  $p(\mathbf{x})$  is then said to satisfy the **ordered Markov property**.
- ▶ By definition of conditional independence:  
 $p(x_i | x_1, \dots, x_{i-1}) = p(x_i | \text{pre}_i) = p(x_i | \pi_i)$
- ▶ With the convention  $\pi_1 = \emptyset$ , we obtain the factorisation

$$p(x_1, \dots, x_d) = \prod_{i=1}^d p(x_i | \pi_i)$$

- ▶ See later: the  $\pi_i$  correspond to the parents of  $x_i$  in graphs.

# From (conditional) independence to factorisation

- ▶ Assume the variables are ordered as  $x_1, \dots, x_d$ , let  $\text{pre}_i = \{x_1, \dots, x_{i-1}\}$  and  $\pi_i \subseteq \text{pre}_i$ .
- ▶ We have seen that

$$\begin{array}{l} \text{if} \quad x_i \perp\!\!\!\perp (\text{pre}_i \setminus \pi_i) \mid \pi_i \text{ for all } i \\ \text{then} \quad p(x_1, \dots, x_d) = \prod_{i=1}^d p(x_i \mid \pi_i) \end{array}$$

- ▶ The chain rule corresponds to the case where  $\pi_i = \text{pre}_i$ .
- ▶ Do we also have the reverse?

$$\begin{array}{l} \text{if} \quad p(x_1, \dots, x_d) = \prod_{i=1}^d p(x_i \mid \pi_i) \quad \text{with } \pi_i \subseteq \text{pre}_i \\ \text{then} \quad x_i \perp\!\!\!\perp (\text{pre}_i \setminus \pi_i) \mid \pi_i \text{ for all } i? \end{array}$$

# From factorisation to (conditional) independence

- ▶ Let us first check whether  $x_d \perp\!\!\!\perp (\text{pre}_d \setminus \pi_d) \mid \pi_d$  holds.
- ▶ We do that by checking whether

$$p(x_d \mid \overbrace{x_1, \dots, x_{d-1}}^{\text{pre}_d}) = p(x_d \mid \pi_d)$$

holds.

- ▶ Since

$$p(x_d \mid x_1, \dots, x_{d-1}) = \frac{p(x_1, \dots, x_d)}{p(x_1, \dots, x_{d-1})}$$

we start with computing  $p(x_1, \dots, x_{d-1})$ .



# From factorisation to (conditional) independence

Assume that the  $x_i$  are ordered as  $x_1, \dots, x_d$  and that  $p(x_1, \dots, x_d) = \prod_{i=1}^d p(x_i | \pi_i)$  with  $\pi_i \subseteq \text{pre}_i$ .

We compute  $p(x_1, \dots, x_{d-1})$  using the sum rule:

$$\begin{aligned} p(x_1, \dots, x_{d-1}) &= \int p(x_1, \dots, x_d) dx_d \\ &= \int \prod_{i=1}^d p(x_i | \pi_i) dx_d \\ &= \int \prod_{i=1}^{d-1} p(x_i | \pi_i) p(x_d | \pi_d) dx_d \quad (x_d \notin \pi_i, i < d) \\ &= \prod_{i=1}^{d-1} p(x_i | \pi_i) \int p(x_d | \pi_d) dx_d \\ &= \prod_{i=1}^{d-1} p(x_i | \pi_i) \end{aligned}$$

# From factorisation to (conditional) independence

Hence:

$$\begin{aligned} p(x_d | x_1, \dots, x_{d-1}) &= \frac{p(x_1, \dots, x_d)}{p(x_1, \dots, x_{d-1})} \\ &= \frac{\prod_{i=1}^d p(x_i | \pi_i)}{\prod_{i=1}^{d-1} p(x_i | \pi_i)} \\ &= p(x_d | \pi_d) \end{aligned}$$

And  $p(x_d | x_1, \dots, x_{d-1}) = p(x_d | \text{pre}_d) = p(x_d | \pi_d)$  means that  $x_d \perp\!\!\!\perp (\text{pre}_d \setminus \pi_d) \mid \pi_d$  as desired.

$p(x_1, \dots, x_{d-1})$  has the same form as  $p(x_1, \dots, x_d)$ : apply same procedure to all  $p(x_1, \dots, x_k)$ , for smaller and smaller  $k \leq d - 1$

Proves that

- (1)  $p(x_1, \dots, x_k) = \prod_{i=1}^k p(x_i | \pi_i)$  and that
- (2) factorisation implies  $x_i \perp\!\!\!\perp (\text{pre}_i \setminus \pi_i) \mid \pi_i$  for all  $i$

# Brief summary

- ▶ Let  $\mathbf{x} = (x_1, \dots, x_d)$  be a  $d$ -dimensional random vector with pdf/pmf  $p(\mathbf{x})$ .
- ▶ Denote the predecessors of  $x_i$  in the ordering by  $\text{pre}(x_i) = \text{pre}_i = \{x_1, \dots, x_{i-1}\}$ , and let  $\pi_i \subseteq \text{pre}_i$ .

$$p(\mathbf{x}) = \prod_{i=1}^d p(x_i | \pi_i) \iff x_i \perp\!\!\!\perp (\text{pre}_i \setminus \pi_i) \mid \pi_i \text{ for all } i$$

- ▶ Equivalence of factorisation and ordered Markov property of the pdf/pmf

# Why does it matter?

- ▶ Denote the predecessors of  $x_i$  in the ordering by  $\text{pre}_i = \{x_1, \dots, x_{i-1}\}$ , and let  $\pi_i \subseteq \text{pre}_i$ .

$$p(\mathbf{x}) = \prod_{i=1}^d p(x_i | \pi_i) \iff x_i \perp\!\!\!\perp (\text{pre}_i \setminus \pi_i) \mid \pi_i \text{ for all } i$$

- ▶ Why does it matter?
  - ▶ Relatively strong result: It holds for sets of pdfs/pmfs and not only single instances
  - ▶ For all members of the set: Fewer numbers are needed for their representation (computational advantage)
  - ▶ Given the independencies, we know what form  $p(\mathbf{x})$  must have (helpful for specifying models)
  - ▶ Increased understanding of the properties of the model (independencies and data generation mechanism)
  - ▶ Visualisation as a graph

# Program

1. Equivalence of factorisation and ordered Markov property
  - Chain rule
  - Ordered Markov property implies factorisation
  - Factorisation implies ordered Markov property
2. Understanding models from their factorisation

# Program

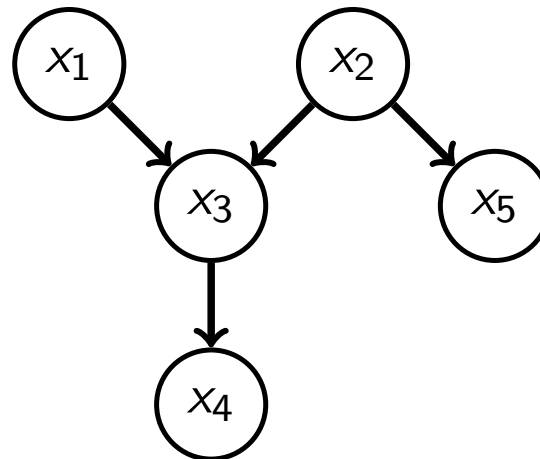
1. Equivalence of factorisation and ordered Markov property
2. Understanding models from their factorisation
  - Visualisation as a directed graph
  - Description of directed graphs and topological orderings

# Visualisation as a directed graph

If  $p(\mathbf{x}) = \prod_{i=1}^d p(x_i | \pi_i)$  with  $\pi_i \subseteq \text{pre}_i$  we can visualise the model as a graph with the random variables  $x_i$  as nodes, and directed edges that point from the  $x_j \in \pi_i$  to the  $x_i$ . This results in a directed acyclic graph (DAG).

Example:

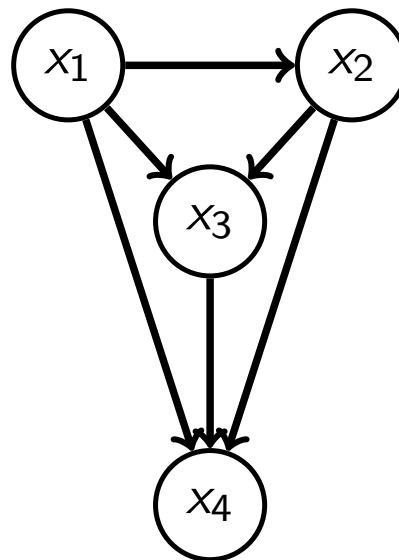
$$p(x_1, x_2, x_3, x_4, x_5) = p(x_1)p(x_2)p(x_3|x_1, x_2)p(x_4|x_3)p(x_5|x_2)$$



# Visualisation as a directed graph

Example:

$$p(x_1, x_2, x_3, x_4) = p(x_1)p(x_2|x_1)p(x_3|x_1, x_2)p(x_4|x_1, x_2, x_3)$$

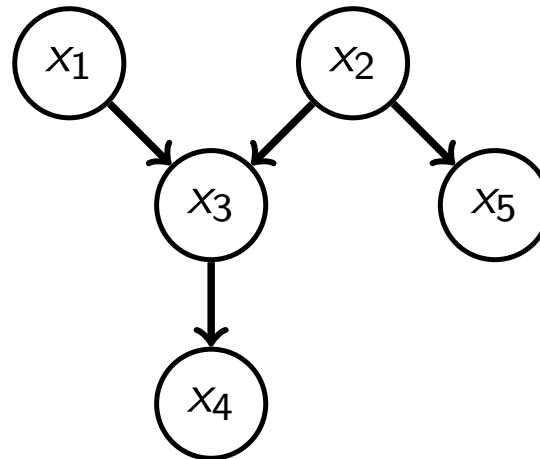


Factorisation obtained by chain rule  $\equiv$  fully connected directed acyclic graph.



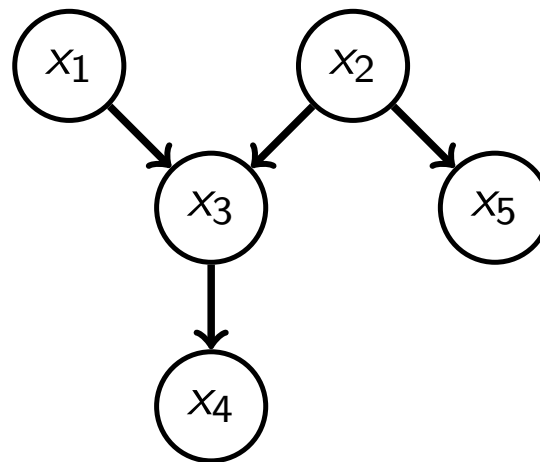
# Graph concepts

- ▶ **Directed graph:** graph where all edges are directed
- ▶ **Directed acyclic graph (DAG):** by following the direction of the arrows you will never visit a node more than once
- ▶  $x_i$  is a **parent** of  $x_j$  if there is a (directed) edge from  $x_i$  to  $x_j$ . The set of parents of  $x_i$  in the graph is denoted by  $\text{pa}(x_i) = \text{pa}_i$ , e.g.  $\text{pa}(x_3) = \text{pa}_3 = \{x_1, x_2\}$ .
- ▶  $x_j$  is a **child** of  $x_i$  if  $x_j \in \text{pa}(x_i)$ , e.g.  $x_3$  and  $x_5$  are children of  $x_2$ .



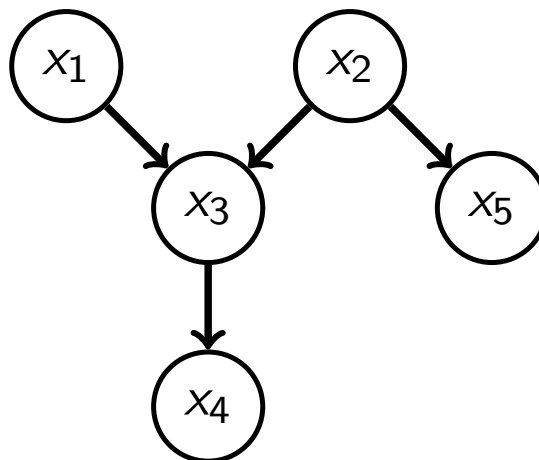
# Graph concepts

- ▶ A **path** or **trail** from  $x_i$  to  $x_j$  is a sequence of distinct connected nodes starting at  $x_i$  and ending at  $x_j$ . The direction of the arrows does *not* matter. For example:  $x_5, x_2, x_3, x_1$  is a trail.
- ▶ A **directed path** is a sequence of connected nodes where we follow the direction of the arrows. For example:  $x_1, x_3, x_4$  is a directed path. But  $x_5, x_2, x_3, x_1$  is not a directed path.



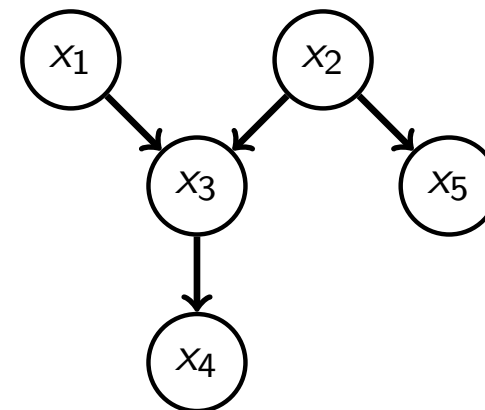
# Graph concepts

- ▶ The **ancestors**  $\text{anc}(x_i)$  of  $x_i$  are all the nodes where a directed path leads to  $x_i$ . For example,  $\text{anc}(x_4) = \{x_1, x_3, x_2\}$ .
- ▶ The **descendants**  $\text{desc}(x_i)$  of  $x_i$  are all the nodes that can be reached on a directed path from  $x_i$ . For example,  $\text{desc}(x_1) = \{x_3, x_4\}$ .  
(Note: sometimes,  $x_i$  is included in the set of ancestors and descendants)
- ▶ The **non-descendants** of  $x_i$  are all the nodes in a graph without  $x_i$  and without the descendants of  $x_i$ . For example,  $\text{nondesc}(x_3) = \{x_1, x_2, x_5\}$



# Graph concepts

- ▶ **Topological ordering:** an ordering  $(x_1, \dots, x_d)$  of some variables  $x_i$  is topological relative to a graph if parents come before their children in the ordering. (whenever there is a directed edge from  $x_i$  to  $x_j$ ,  $x_i$  occurs prior to  $x_j$  in the ordering.)
- ▶ There is always at least one such ordering for DAGs.
- ▶ For a pdf  $p(\mathbf{x})$ , assume you order the random variables  $x_i$  in some manner and compute the corresponding factorisation, e.g.  $p(\mathbf{x}) = p(x_1)p(x_2)p(x_3|x_1, x_2)p(x_4|x_3)p(x_5|x_2)$
- ▶ When you visualise the factorised pdf as a graph, the graph is always such that the ordering used for the factorisation is topological to it.
- ▶ The  $\pi_i$  in the factorisation are equal to the parents  $pa_i$  in the graph. We may call both sets the “parents” of  $x_i$ .



# Program recap

1. Equivalence of factorisation and ordered Markov property
  - Chain rule
  - Ordered Markov property implies factorisation
  - Factorisation implies ordered Markov property
2. Understanding models from their factorisation
  - Visualisation as a directed graph
  - Description of directed graphs and topological orderings