

# 732A96/TDDE15 Advanced Machine Learning

## Graphical Models

Jose M. Peña  
IDA, Linköping University, Sweden

Lecture 1: Bayesian and Markov Networks

# Contents

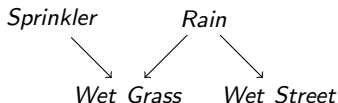
- ▶ Causal Structures
- ▶ Bayesian Networks
  - ▶ Definition
  - ▶ Causal Inference
  - ▶ Probabilistic Inference
- ▶ Markov Networks
  - ▶ Definition
  - ▶ Probabilistic Inference

# Literature

- ▶ Main source
  - ▶ Bishop, C. M. *Pattern Recognition and Machine Learning*. Springer, 2006. Chapter 8.
- ▶ Additional source
  - ▶ Koski, T. J. T. and Noble, J. M. A Review of Bayesian Networks and Structure Learning. *Mathematica Applicanda* 40, 51-103, 2012.

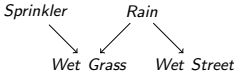
## Causal Structures

- ▶ Assume that we want to represent the causal relations between a set of random variables, e.g. the variables may represent the state of the components of a system.
- ▶ A natural and intuitive representation consists of a **graph** where the nodes are the random variables, and the edges are the causal relations between the variables. We call such a graph a **causal structure**.



- ▶ **Exercise.** Produce a causal structure for the domain *Temperature*, *Ice cream sales* and *Soda sales*.
- ▶ **Exercise.** Produce a causal structure for Boyle's law, which relates the pressure and volume of a gas as  $Pressure \cdot Volume = constant$  if the temperature and amount of gas remain unchanged within a closed system.

# Bayesian Networks: Definition

DAG	Parameter values for the conditional probability distributions
 <pre> graph TD     Sprinkler --&gt; WetGrass[Wet Grass]     Rain --&gt; WetGrass     Rain --&gt; WetStreet[Wet Street]         </pre>	$q(s) = (0.3, 0.7) = (\theta_{s_0}, \theta_{s_1})$ $q(r) = (0.5, 0.5) = (\theta_{r_0}, \theta_{r_1})$ $q(wg r_0, s_0) = (0.1, 0.9) = (\theta_{wg_0 r_0, s_0}, \theta_{wg_1 r_0, s_0})$ $q(wg r_0, s_1) = (0.7, 0.3) = (\theta_{wg_0 r_0, s_1}, \theta_{wg_1 r_0, s_1})$ $q(wg r_1, s_0) = (0.8, 0.2) = (\theta_{wg_0 r_1, s_0}, \theta_{wg_1 r_1, s_0})$ $q(wg r_1, s_1) = (0.9, 0.1) = (\theta_{wg_0 r_1, s_1}, \theta_{wg_1 r_1, s_1})$ $q(ws r_0) = (0.1, 0.9) = (\theta_{ws_0 r_0}, \theta_{ws_1 r_0})$ $q(ws r_1) = (0.7, 0.3) = (\theta_{ws_0 r_1}, \theta_{ws_1 r_1})$ $p(s, r, wg, ws) = q(s)q(r)q(wg s, r)q(ws r)$

- ▶ A **Bayesian network (BN)** over a finite set of **discrete** random variables  $X = X_{1:n} = \{X_1, \dots, X_n\}$  consists of
  - ▶ a directed acyclic graph (DAG)  $G$  whose nodes are the elements in  $X$ , and
  - ▶ parameter values  $\theta$  specifying probability distributions  $q(x_i|pa_i)$ , where  $Pa_i$  are the parents of  $X_i$  in  $G$ , i.e. the nodes with an edge into  $X_i$ .
- ▶ The BN represents a **causal** model of the system.
- ▶ And also a **probabilistic** model of the system as  $p(x) = \prod_i q(x_i|pa_i)$ .

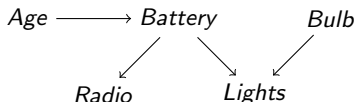
## Bayesian Networks: Definition

- ▶ We now show that  $p(x) = \prod_i q(x_i|pa_i)$  is a probability distribution.
- ▶ Clearly,  $0 \leq \prod_i q(x_i|pa_i) \leq 1$ .
- ▶ Assume without loss of generality that  $Pa_i \subseteq X_{1:i-1}$  for all  $i$ . Then
$$\sum_x \prod_i q(x_i|pa_i) = \sum_{x_1} [q(x_1) \dots \sum_{x_{n-1}} [q(x_{n-1}|pa_{n-1}) \sum_{x_n} q(x_n|pa_n)] \dots] = 1$$
- ▶ Moreover,  $p(x_j|pa_j) = q(x_j|pa_j)$ . To see it, note that

$$\begin{aligned} p(x_j|pa_j) &= \frac{p(x_j, pa_j)}{p(pa_j)} = \frac{\sum_{x \setminus \{x_j, pa_j\}} \prod_i q(x_i|pa_i)}{\sum_{x \setminus pa_j} \prod_i q(x_i|pa_i)} \\ &= \frac{\sum_{x_{1:j} \setminus \{x_j, pa_j\}} \prod_{i \leq j} q(x_i|pa_i)}{\sum_{x_{1:j} \setminus pa_j} \prod_{i \leq j} q(x_i|pa_i)} = q(x_j|pa_j) \end{aligned}$$

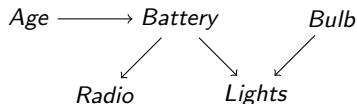
## Bayesian Networks: Separation

- ▶ We now show that many of the independencies in  $p$  can be read off  $G$  without numerical calculations. Consider the following DAG.



- ▶ **Chain:**  $Age \rightarrow Battery \rightarrow Radio$ 
  - ▶  $Age \not\perp Radio | \emptyset$
  - ▶  $Age \perp Radio | Battery$
- ▶ **Fork:**  $Radio \leftarrow Battery \rightarrow Lights$ 
  - ▶  $Radio \not\perp Lights | \emptyset$
  - ▶  $Radio \perp Lights | Battery$
- ▶ **Collider:**  $Battery \rightarrow Lights \leftarrow Bulb$ 
  - ▶  $Battery \perp Bulb | \emptyset$
  - ▶  $Battery \not\perp Bulb | Lights$
- ▶ **Chain + collider:**  $Age \rightarrow Battery \rightarrow Lights \leftarrow Bulb$ 
  - ▶  $Age \perp Bulb | \emptyset$
  - ▶  $Age \not\perp Bulb | Lights$
  - ▶  $Age \perp Bulb | Lights, Battery$

## Bayesian Networks: Separation

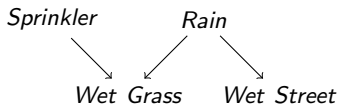


- ▶ A path in  $G$  is a sequence of distinct and adjacent nodes, i.e. the direction of the edge is irrelevant. A node  $B$  is a descendant of a node  $A$  in  $G$  if there is a path  $A \rightarrow \dots \rightarrow B$ .
  - ▶ E.g.,  $Age \rightarrow Battery \rightarrow Lights \leftarrow Bulb$  is a path.
  - ▶ E.g.,  $Lights$  is a descendant of  $Age$ .
- ▶ Let  $\rho$  be a path in  $G$  between the nodes  $\alpha$  and  $\beta$ .
- ▶ A node  $B$  in  $\rho$  is a **collider** when  $A \rightarrow B \leftarrow C$  is a subpath of  $\rho$ .
  - ▶ E.g.,  $Lights$  is a collider in the path  $Age \rightarrow Battery \rightarrow Lights \leftarrow Bulb$ .
- ▶ Moreover,  $\rho$  is  **$Z$ -open** with  $Z \subseteq X \setminus \{\alpha, \beta\}$  when
  - ▶ no non-collider in  $\rho$  is in  $Z$ , and
  - ▶ every collider in  $\rho$  is in  $Z$  or has a descendant in  $Z$ .
  - ▶ E.g., the path  $Age \rightarrow Battery \rightarrow Lights \leftarrow Bulb$  is  $Z$ -open with  $Z = \{Lights\}$ .
- ▶ Let  $U$ ,  $V$  and  $Z$  be three disjoint subsets of  $X$ . Then,  $U$  and  $V$  are **separated** given  $Z$  in  $G$  (i.e.  $U \perp_G V | Z$ ) when there is no  $Z$ -open path in  $G$  between a node in  $U$  and a node in  $V$ .
  - ▶ E.g.,  $Age, Battery \perp_G Bulb | \emptyset$ .



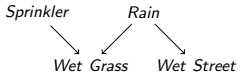

## Bayesian Networks: Separation

- ▶ The separation criterion is **sound**, i.e. if  $U \perp_G V|Z$  then  $U \perp_p V|Z$ .
- ▶ For instance,  $S \perp_p R$ ,  $S \not\perp_p R|WG$ ,  $S \not\perp_p WS|WG$ ,  $S \perp_p WS|WG, R$  follow from the DAG



- ▶ Note that we read independencies from  $G$ , never dependencies.
- ▶ **Exercise.** Prove that  $A \perp_p B|C$  for the DAGs  $A \rightarrow C \rightarrow B$ ,  $A \leftarrow C \rightarrow B$  and  $A \leftarrow C \leftarrow B$ , i.e. prove that  $p(a, b|c) = p(a|c)p(b|c)$ .
- ▶ **Exercise.** Prove that  $A \perp_p B|\emptyset$  for the DAG  $A \rightarrow C \leftarrow B$ , i.e. prove that  $p(a, b) = p(a)p(b)$ .
- ▶ **Exercise.** Find the minimal set of nodes that separates a given node from the rest. This set is called the Markov blanket of the given node.
- ▶ **Exercise.** How many free parameters do we have in the wet grass BN ? How many do we have if we specify the distribution without the assistance of a BN, i.e. as a table ?

# Bayesian Networks: Causal Inference

Original	After $do(r_1)$
 <p> <math>q(s) = (0.3, 0.7)</math>  <math>q(r) = (0.5, 0.5)</math>  <math>q(wg r_0, s_0) = (0.1, 0.9)</math>  <math>q(wg r_0, s_1) = (0.7, 0.3)</math>  <math>q(wg r_1, s_0) = (0.8, 0.2)</math>  <math>q(wg r_1, s_1) = (0.9, 0.1)</math>  <math>q(ws r_0) = (0.1, 0.9)</math>  <math>q(ws r_1) = (0.7, 0.3)</math>  <math>p(s, r, wg, ws) = q(s)q(r)q(wg s, r)q(ws r)</math> </p>	 <p> <math>q(s) = (0.3, 0.7)</math>  <math>q(wg s_0) = (0.8, 0.2)</math>  <math>q(wg s_1) = (0.9, 0.1)</math>  <math>q(ws) = (0.7, 0.3)</math>  <math>p(s, wg, ws) = q(s)q(wg s)q(ws)</math> </p>

- ▶ What would be the state of the system if a random variable  $X_j$  is **forced** to take the state  $x_j$  ?
  - ▶ Remove  $X_j$  and all the edges from and to  $X_j$  from  $G$ .
  - ▶ Remove  $q(x_j|pa_j)$ .
  - ▶ If  $X_j \in Pa_i$ , then replace  $q(x_i|pa_i)$  with  $q(x_i|pa_i \setminus x_j, x_j)$
  - ▶ Set  $p(x \setminus x_j|do(x_j)) = \prod_i q(x_i|pa_i)$ .
- ▶ So, the result of  $do(x)$  on a BN is a **BN**. No more on causality in this course.

## Bayesian Networks: Probabilistic Inference

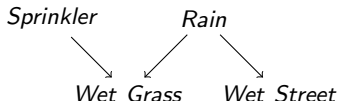
- ▶ What is the state of a random variable  $X_k$  if a random variable  $X_i$  is **observed** to be in the state  $x_i$  ?

$$p(x_k|x_i) = \frac{p(x_k, x_i)}{p(x_i)} = \frac{\sum_{x \setminus \{x_i, x_k\}} p(x)}{\sum_{x \setminus x_i} p(x)}$$

- ▶ For instance,

$$p(ws|s) = \frac{\sum_{r, wg} p(r, wg, ws, s)}{\sum_{r, wg, ws} p(r, wg, ws, s)} = \frac{\sum_{r, wg} q(s)q(r)q(wg|s, r)q(ws|r)}{\sum_{r, wg, ws} q(s)q(r)q(wg|s, r)q(ws|r)}$$


for the DAG



- ▶ Answering questions like the one above can be computationally hard.
- ▶ A BN is an efficient (because it uses the independences encoded) formalism to compute a posterior probability distribution from a prior probability distribution in the light of observations, hence the name. More on probabilistic inference in Lecture 2.

## Markov Networks: Definition

- ▶ A BN represents asymmetric (causal) relations, whereas a Markov network represents **symmetric** relations, e.g. physical laws.

UG	Potentials assuming binary random variables
 <pre> graph LR     A --- B     A --- C     B --- D     C --- D         </pre>	$\varphi(a, b, c) = (0, 0, 0, 0, 1, 1, 1, 1)$ $\varphi(b, c, d) = (1, 2, 3, 4, 5, 6, 7, 8)$  $p(a, b, c, d) = \varphi(a, b, c)\varphi(b, c, d)/Z$ with $Z = \sum_{a,b,c,d} \varphi(a, b, c)\varphi(b, c, d)$

- ▶ A **Markov network (MN)** over  $X$  consists of
  - ▶ an undirected graph (UG)  $G$  whose nodes are the elements in  $X$ , and
  - ▶ a set of non-negative functions  $\varphi(k)$  over the cliques  $Cl(G)$  of  $G$ , i.e. the maximal complete sets of nodes in  $G$ . The functions are called potentials. They represent **compatibility** relations between the random variables in the cliques.
- ▶ The MN represents a **probabilistic** model of the system, namely

$$p(x) = \frac{1}{Z} \prod_{K \in Cl(G)} \varphi(k)$$

where  $Z$  is a normalization constant, i.e.

$$Z = \sum_x \prod_{K \in Cl(G)} \varphi(k)$$

- ▶ Clearly,  $p(x)$  is a probability distribution.

## Markov Networks: Separation

- ▶ We now show that many of the independencies in  $p$  can be read off  $G$  without numerical calculations.
- ▶ A path  $\rho$  in  $G$  between two nodes  $\alpha$  and  $\beta$  is  **$Z$ -open** with  $Z \subseteq X \setminus \{\alpha, \beta\}$  when no node in  $\rho$  is in  $Z$ .
- ▶ Let  $U$ ,  $V$  and  $Z$  be three disjoint subsets of  $X$ . Then,  $U$  and  $V$  are **separated** given  $Z$  in  $G$  (i.e.  $U \perp_G V | Z$ ) when there is no  $Z$ -open path in  $G$  between a node in  $U$  and a node in  $V$ .
- ▶ The separation criterion is **sound**, i.e. if  $U \perp_G V | Z$  then  $U \perp_p V | Z$ .

## Markov Networks: Separation

- ▶ **Exercise.** Prove that  $A \perp_p B | C$  for the UG  $A - C - B$ , i.e. prove that  $p(a, b | c) = f(a, c)g(b, c)$  for some functions  $f$  and  $g$ .
- ▶ **Exercise.** Find the minimal set of nodes that separates a given node from the rest. This set is called the Markov blanket of the given node.
- ▶ **Exercise.** How many free parameters do we have in the ABCD MN ? How many do we have if we specify the distribution without the assistance of a MN ? How many if the variables have three states ?

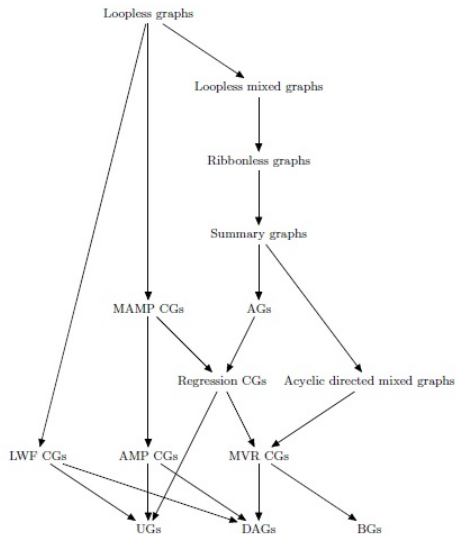
# Markov Networks: Probabilistic Inference

- ▶ What is the state of a random variable  $A$  if a random variable  $B$  is **observed** to be in the state  $b$  ?

$$p(a|b) = \frac{\sum_{c,d} \varphi(a, b, c) \varphi(b, c, d) / Z}{\sum_{a,c,d} \varphi(a, b, c) \varphi(b, c, d) / Z}$$

- ▶ Answering questions like the one above can be computationally hard.
- ▶ A MN is an efficient (because it uses the independences encoded) formalism to answer such questions. More on probabilistic inference in Lecture 2.

# Families of Graphical Models





# Contents

- ▶ Causal Structures
- ▶ Bayesian Networks
  - ▶ Definition
  - ▶ Causal Inference
  - ▶ Probabilistic Inference
- ▶ Markov Networks
  - ▶ Definition
  - ▶ Probabilistic Inference

Thank you