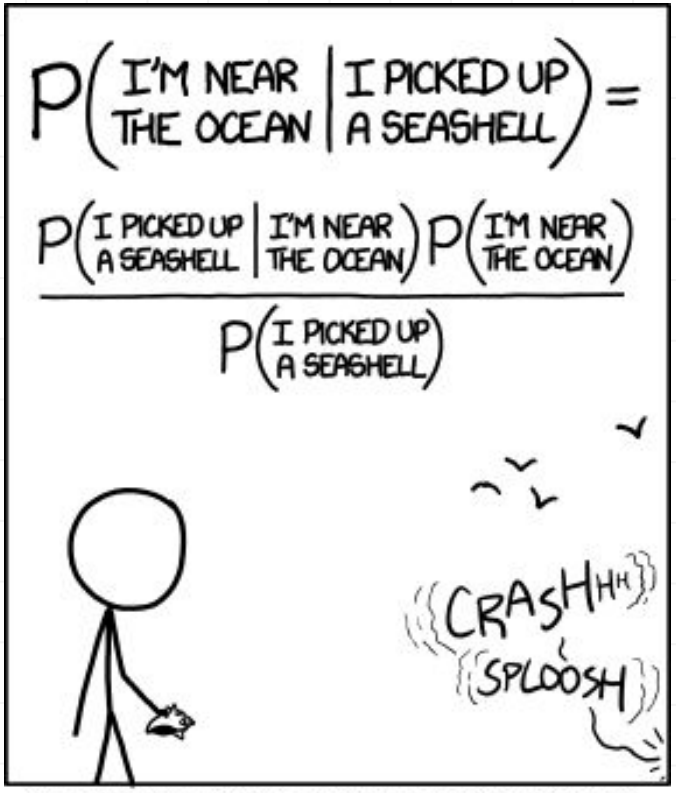
https://xkcd.com/1236/



STATISTICALLY SPEAKING, IF YOU PICK UP A SEASHELL AND *DON'T* HOLD IT TO YOUR EAR, YOU CAN PROBABLY HEAR THE OCEAN.

Graphical models

[Bishop 8.1 and 8.2]

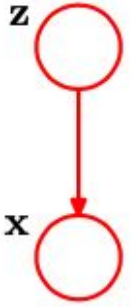
What? And Why?

What is graphical models / Bayesian network

Plate notation

Conditional independence

Fig 9.4

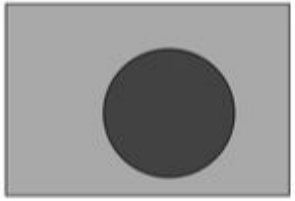


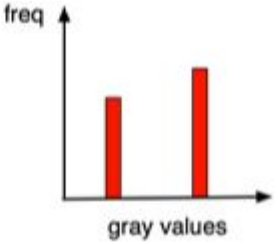
Quiz 2 remark

Everything comes from the sum rule and the product rule … but graphical models provide a rich set of modeling language and inference procedures.

Motivating scenario - computer vision

- Image Segmentation
  Cluster the gray value representation of an image

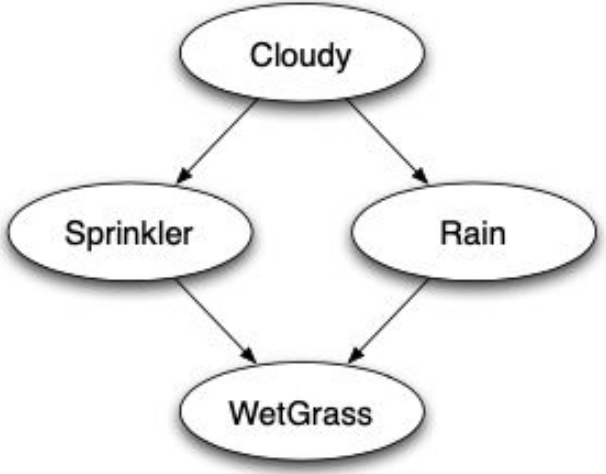




- Neighbourhood information lost
  Need to use the structure of the image.

Motivation - encoding (in)dependence

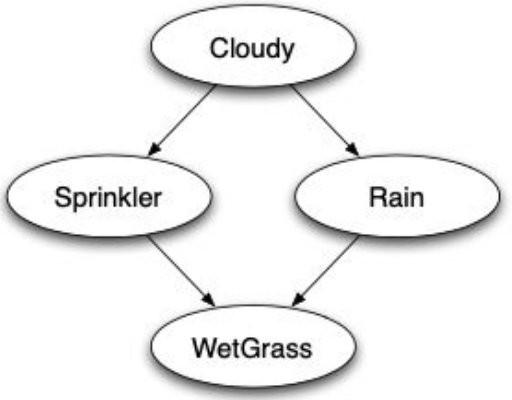
Why is the grass wet?



Introduce four Boolean variables :

$C(loudy), S(prinkler), R(ain), W(etGrass) \in \{F(alse), T(rue)\}.$

Motivation - encoding (in)dependence



Model the conditional probabilities

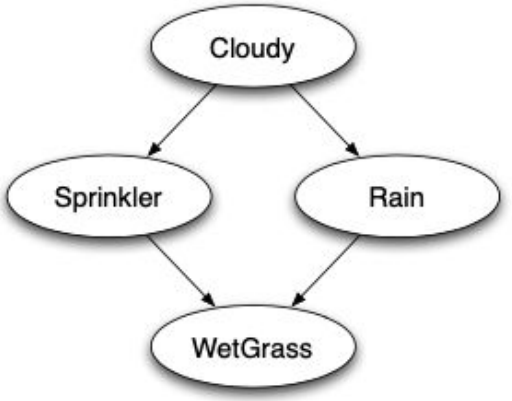
p(C = F)	p(C = T)
0.2	0.8

C	p(S = F)	p(S = T)
F	0.5	0.5
T	0.9	0.1

C	p(R = F)	p(R = T)
F	0.8	0.2
T	0.2	0.8

S R	p(W = F)	p(W = T)
F F	1.0	0.0
T F	0.1	0.9
F T	0.1	0.9
T T	0.01	0.99

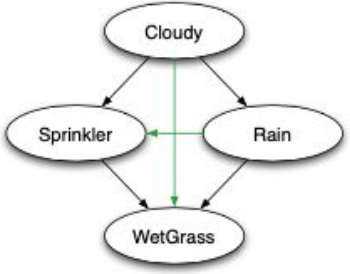
Motivation - encoding (in)dependence



If everything depends on everything

C S R W	p(C, S, R, W)
F F F F	...
F F F T	...
...	...
T T T F	...
T T T T	...

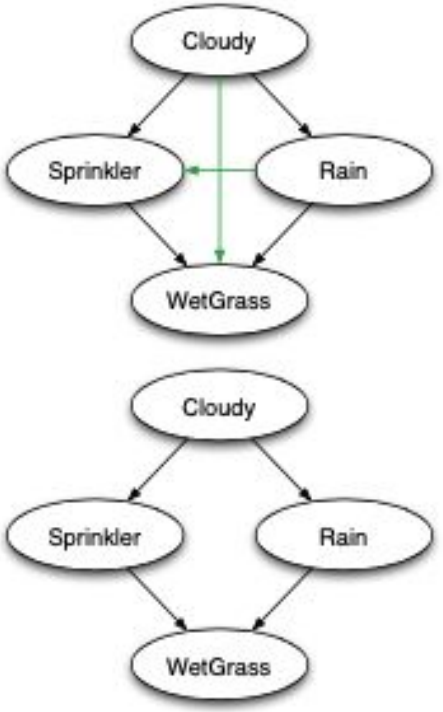
$$
\begin{aligned}
p(W,S,R,C) &= p(W \mid S,R,C)\, p(S,R,C) \\
&= p(W \mid S,R,C)\, p(S \mid R,C)\, p(R,C) \\
&= p(W \mid S,R,C)\, p(S \mid R,C)\, p(R \mid C)\, p(C)
\end{aligned}
$$



Motivation - encoding (in)dependence

$$p(W) = \sum_{S,R,C} p(W \mid S,R,C)\, p(S \mid R,C)\, p(R \mid C)\, p(C)$$

$$p(W) = \sum_{S,R} p(W \mid S,R) \sum_{C} p(S \mid C)\, p(R \mid C)\, p(C)$$



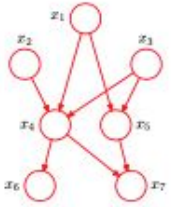
Probabilistic graphical model - terminologies

Nodes: random variables

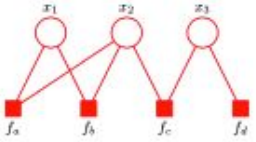
Edges: probabilistic relationships

Graph: captures “structures” in the joint distribution

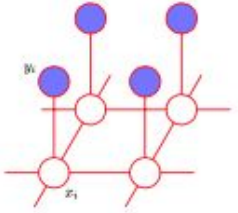
- Directed Graph : Bayesian Network (also called Directed Graphical Model) expressing causal relationship between variables
- Undirected Graph : Markov Random Field expressing soft constraints between variables
- Factor Graph : convenient for solving inference problems (derived from Bayesian Networks or Markov Random Fields).



Bayesian Network



Factor Graph



Markov Random Field

Inference (Ch 8.4) - compute the posterior distributions of one or more subsets of other nodes, given values of a subset of nodes (in a known graphical model with known parameters)

Learning - estimate the parameters (conditional probabilities) for a given graphical model, from a dataset of observed values.

Probabilistic graphical model - overarching goals

Nodes: random variables

Edges: probabilistic relationships

Graph: captures “structures” in the joint distribution

1. They provide a simple way to visualize the structure of a probabilistic model and can be used to design and motivate new models.

2. Insights into the properties of the model, including conditional independence properties, can be obtained by inspection of the graph.

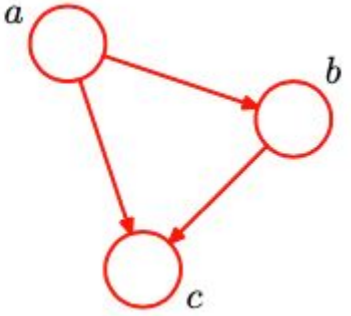
3. Complex computations, required to perform inference and learning in sophisticated models, can be expressed in terms of graphical manipulations, in which underlying mathematical expressions are carried along implicitly.

Bayesian networks

$$p(a, b, c) = p(c|a, b)p(a, b). \quad (8.1)$$

$$p(a, b, c) = p(c|a, b)p(b|a)p(a). \quad (8.2)$$

Figure 8.1 A directed graphical model representing the joint probability distribution over three variables a, b, and c, corresponding to the decomposition on the right-hand side of (8.2).



LHS of (8.2) is symmetric w.r.t *a, b, c*, but RHS is not.

➜ Fig 8.2 implicitly chose an ordering

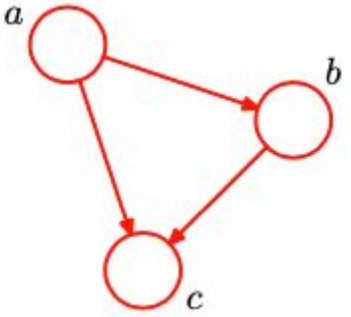
➜ possible to have different ordering, different decomposition, different graphical models

Bayesian networks for any distribution

$$p(a, b, c) = p(c|a, b)p(a, b). \tag{8.1}$$

$$p(a, b, c) = p(c|a, b)p(b|a)p(a). \tag{8.2}$$

Figure 8.1 A directed graphical model representing the joint probability distribution over three variables a, b, and c, corresponding to the decomposition on the right-hand side of (8.2).



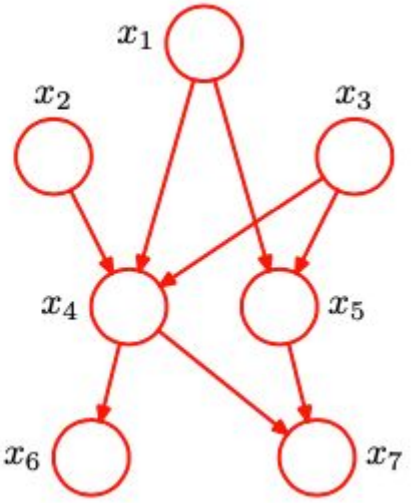
$$p(x_1, \ldots, x_K) = p(x_K|x_1, \ldots, x_{K-1}) \ldots p(x_2|x_1)p(x_1). \tag{8.3}$$

This decomposition holds for any distribution over (x_1, ... x_K).
Graph is “fully connected” -- there is a link (in either direction) between any pair of nodes.

Factoring joint distributions ➔ Graph

$$p(x_1)p(x_2)p(x_3)p(x_4|x_1,x_2,x_3)p(x_5|x_1,x_3)p(x_6|x_4)p(x_7|x_4,x_5). \quad (8.4)$$

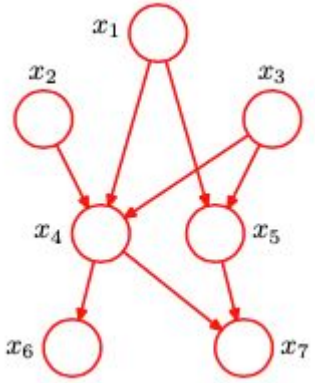
Figure 8.2 Example of a directed acyclic graph describing the joint distribution over variables $x_1,\ldots,x_7$. The corresponding decomposition of the joint distribution is given by (8.4).



... it is the *absence* of links in the graph that conveys interesting information about the properties of the class of distributions that the graph represents.

1. Draw a node for each conditional distribution associated with a random variable.
2. Draw an edge from each conditional distribution associated with a random variable to all other conditional distribution which are conditioned on this variable.

Graph ➔ joint distributions

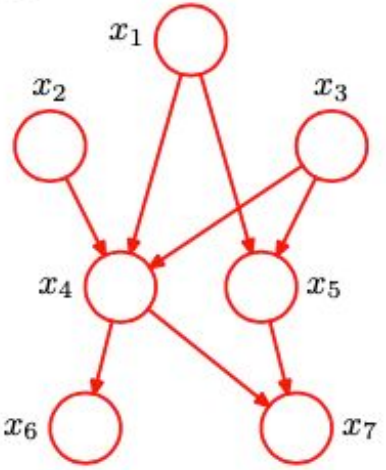


Can we get the expression from the graph?

1. Write a product of probability distributions, one for each associated random variable. ↔ Draw a node for each conditional distribution associated with a random variable.
2. Add all random variables associated with parent nodes to the list of conditioning variables. ↔ Draw an edge from each conditional distribution associated with a random variable to all other conditional distribution which are conditioned on this variable.

$$p(x_1)p(x_2)\,p(x_3)\,p(x_4|x_1,x_2,x_3)\,p(x_5|x_1,x_3)\,p(x_6|x_4)\,p(x_7|x_4,x_5)$$

- Restriction : Graph must be a directed acyclic graph (DAG).
- There are no closed paths in the graph when moving along the directed edges.
- Or equivalently: There exists an ordering of the nodes such that there are no edges that go from any node to any lower numbered node.



- Extension: Can also have sets of variables, or vectors at a node.

General Bayes nets

In general, the joint distribution is given by:

$$p(\mathbf{x}) = \prod_{k=1}^{K} p(x_k|\mathrm{pa}_k) \qquad (8.5)$$

- What does it mean if $\mathrm{pa}(x_k) \neq x_1, x_2, \ldots, x_{k-1}$ in the above equation?
  - Sparse graphical model.
  - $p(\mathbf{x})$ no longer general.
  - Assumption, e.g. $p(W|C,S,R) = p(W|S,R)$.
  - **Conditional independence**.

General Bayes nets

In general, the joint distribution is given by:

$$p(\mathbf{x}) = \prod_{k=1}^{K} p(x_k|\mathrm{pa}_k) \qquad (8.5)$$

Is $p(\mathbf{x})$ normalised, $\sum_{\mathbf{x}} p(\mathbf{x}) = 1$?

- As graph is DAG, there always exists a node with no outgoing edges, say x_i.

$$\sum_{\mathbf{x}} p(\mathbf{x}) = \sum_{x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_K} \prod_{\substack{k=1 \\ k\neq i}}^{K} p(x_k \mid \mathrm{pa}(x_k)) \underbrace{\sum_{x_i} p(x_i \mid \mathrm{pa}(x_i))}_{=1}$$

because $\sum_{x_i} p(x_i \mid \mathrm{pa}(x_i)) = \sum_{x_i} \frac{p(x_i,\mathrm{pa}(x_i))}{p(\mathrm{pa}(x_i))} = \frac{p(\mathrm{pa}(x_i))}{p(\mathrm{pa}(x_i))} = 1$

- Repeat, until no node left.

Plate notation

- Bayesian polynomial regression : observed inputs $\mathbf{x}$, observed targets $\mathbf{t}$, noise variance σ^2, hyperparameter α controlling the priors for $\mathbf{w}$.
- Focusing on $\mathbf{t}$ and $\mathbf{w}$ only

$$p(\mathbf{t}, \mathbf{w}) = p(\mathbf{w}) \prod_{n=1}^{N} p(t_n \mid \mathbf{w})$$

Figure 8.3 Directed graphical model representing the joint distribution (8.6) corresponding to the Bayesian polynomial regression model introduced in Section 1.2.6.

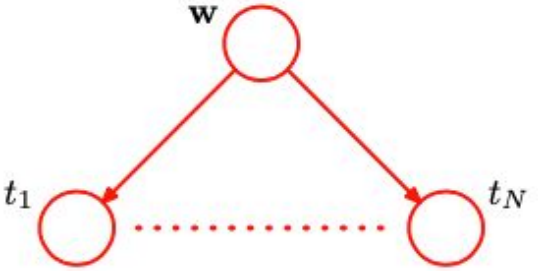
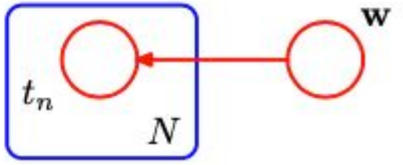


Figure 8.4 An alternative, more compact, representation of the graph shown in Figure 8.3 in which we have introduced a *plate* (the box labelled N) that represents N nodes of which only a single example t_n is shown explicitly.

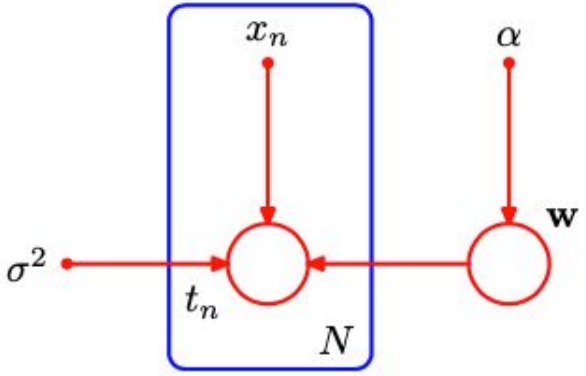


Polynomial regression: more variables

$$p(\mathbf{t}, \mathbf{w} \mid \mathbf{x}, \alpha, \sigma^2) = p(\mathbf{w} \mid \alpha) \prod_{n=1}^{N} p(t_n \mid \mathbf{w}, x_n, \sigma^2)$$

Random variables = open circles

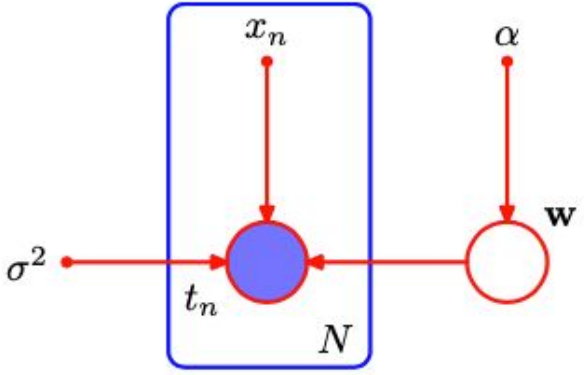
Deterministic variables = smaller solid circles



Random variables

- Observed random variables, e.g. $\mathbf{t}$
- Unobserved random variables, e.g. $\mathbf{w}$, (latent random variables, hidden random variables)

Shade the observed random variables in the graphical model.



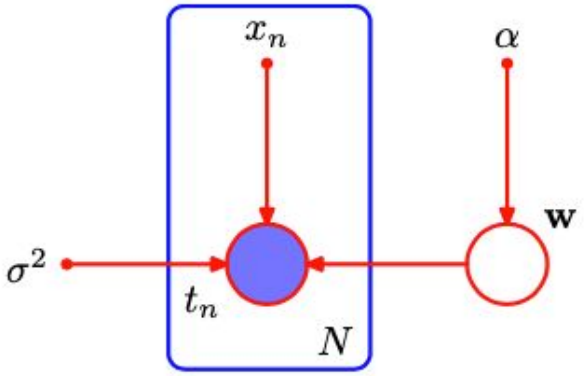
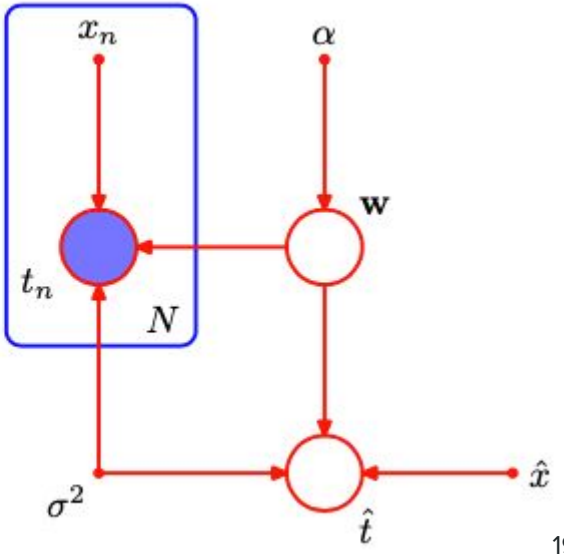


Figure 8.7 The polynomial regression model, corresponding to Figure 8.6, showing also a new input value $\widehat{x}$ together with the corresponding model prediction $\widehat{t}$.



Generative models

Bayes net allows us to draw samples from a (joint) distribution.

Ancestral sampling: $$p(\mathbf{x}) = \prod_{k=1}^{K} p(x_k|\mathrm{pa}_k) \qquad (8.5)$$

- Work though each node *in order*
- Draw lower-number nodes (parents)
- Draw x_k with values of its parents fixed
- To obtain samples from marginals, retain the variables in question and disregard the rest

Generative models

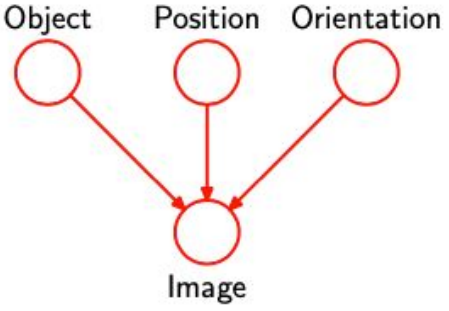
In a typical application

- Higher-numbered nodes: terminal nodes, observed
- Lower-numbered nodes: latent variables, unobserved
  ➔ to allow a complex distribution over the observed to factor into simpler conditional distributions

BN graphs represent causal processes.
Polynomial regression example isn't a generative model (but can be made so with more complexity)

Latent variables need not have physical interpretation (can be introduced for representational/computational reasons)

Figure 8.8 A graphical model representing the process by which images of objects are created, in which the identity of an object (a discrete variable) and the position and orientation of that object (continuous variables) have independent prior probabilities. The image (a vector of pixel intensities) has a probability distribution that is dependent on the identity of the object as well as on its position and orientation.



Number of parameters for discrete variables

Figure 8.9 (a) This fully-connected graph describes a general distribution over two K-state discrete variables having a total of $K^2 - 1$ parameters. (b) By dropping the link between the nodes, the number of parameters is reduced to $2(K - 1)$.

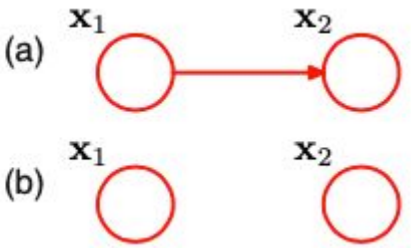


Figure 8.10 This chain of M discrete nodes, each having K states, requires the specification of $K - 1 + (M - 1)K(K - 1)$ parameters, which grows linearly with the length M of the chain. In contrast, a fully connected graph of M nodes would have $K^M - 1$ parameters, which grows exponentially with M.

Read: 8.1.3 in Bishop

Graphical models

What? And Why?

What is graphical models / Bayesian network

Plate notation

Conditional independence

Definition (Conditional Independence)

If for three random variables a, b, and c the following holds

$$p(a \mid b, c) = p(a \mid c)$$

then a is conditionally independent of b given c.
Notation : $a \perp\!\!\!\perp b \mid c$.

- The above equation must hold for all possible values of c.
- Consequence :

$$\begin{aligned} p(a, b \mid c) &= p(a \mid b, c)\, p(b \mid c) \\ &= p(a \mid c)\, p(b \mid c) \end{aligned}$$

- Conditional independence simplifies
  - the structure of the model
  - the computations needed to perform interference/learning.
- NB: conditional **de**pendence is a related but different concept we are not concerned with in this course.

Rules for Conditional Independence

Symmetry : $X \perp\!\!\!\perp Y \,|\, Z \Longrightarrow Y \perp\!\!\!\perp X \,|\, Z$

Decomposition : $Y, W \perp\!\!\!\perp X \,|\, Z \Longrightarrow Y \perp\!\!\!\perp X \,|\, Z$ and $W \perp\!\!\!\perp X \,|\, Z$

Weak Union : $X \perp\!\!\!\perp Y, W \,|\, Z \Longrightarrow X \perp\!\!\!\perp Y \,|\, Z, W$

Contraction : $X \perp\!\!\!\perp W \,|\, Z, Y$

and $X \perp\!\!\!\perp Y \,|\, Z \Longrightarrow X \perp\!\!\!\perp W, Y \,|\, Z$

Intersection : $X \perp\!\!\!\perp Y \,|\, Z, W$

and $X \perp\!\!\!\perp W \,|\, Z, Y \Longrightarrow X \perp\!\!\!\perp Y, W \,|\, Z$

Note: Intersection is only valid for $p(X), p(Y), p(Z), p(W) > 0$.

Three-node graphs

Goal: infer conditional independence of *a* and *b*

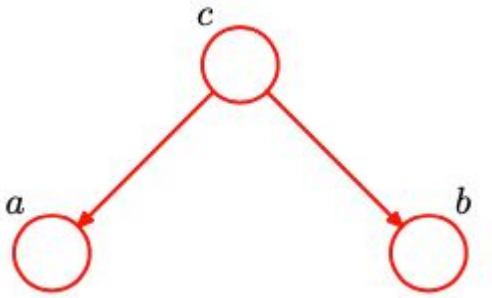
Larger goal: spot general subgraph patterns so as to reason about conditional independence in general.

$$p(a,b,c) = p(a\,|\,c)\,p(b\,|\,c)\,p(c)$$

Marginalise both sides over c

$$p(a,b) = \sum_c p(a\,|\,c)\,p(b\,|\,c)\,p(c) \neq p(a)\,p(b).$$

Does not hold : $a \perp\!\!\!\perp b\,|\,\emptyset$ (where $\emptyset$ is the empty set).

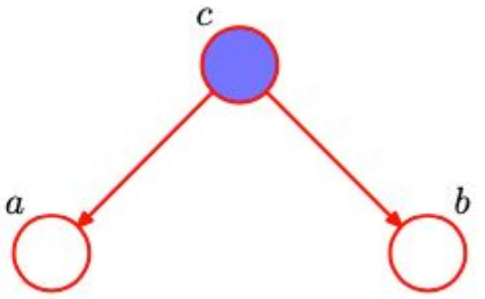


Now condition on c.

$$p(a, b \mid c) = \frac{p(a, b, c)}{p(c)} = p(a \mid c)\, p(b \mid c)$$

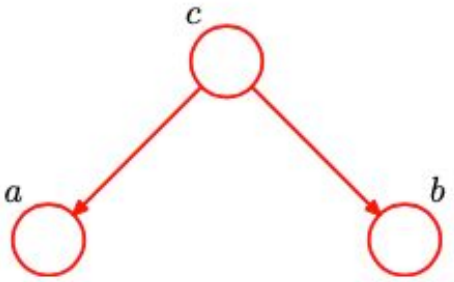
Therefore $a \perp\!\!\!\perp b \mid c$.

Figure 8.16 As in Figure 8.15 but where we have conditioned on the value of variable c.

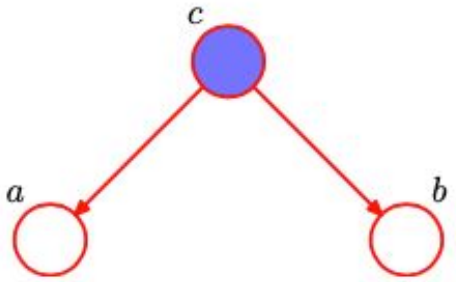


Graphical interpretation

- In both graphical models there is a path from a to b.
- The node c is called tail-to-tail (TT) with respect to this path because the node c is connected to the tails of the arrows in the path.
- The presence of the TT-node c in the path left renders a dependent on b (and b dependent on a).
- Conditioning on c blocks the path from a to b and causes a and b to become conditionally independent on c.



Not $a \perp\!\!\!\perp b \mid \emptyset$



$a \perp\!\!\!\perp b \mid c$

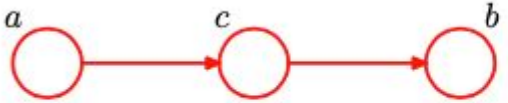
$$p(a,b,c) = p(a)\,p(c\,|\,a)\,p(b\,|\,c)$$

Head-Tail (HT) pattern

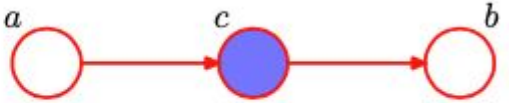
Marginalise over c to test for independence.

$$p(a,b) = p(a)\sum_c p(c\,|\,a)\,p(b\,|\,c) = p(a)\,p(b\,|\,a) \neq p(a)\,p(b)$$

Does not hold : $a \perp\!\!\!\perp b\,|\,\emptyset$.



Not $a \perp\!\!\!\perp b\,|\,\emptyset$



$a \perp\!\!\!\perp b\,|\,c$

Now condition on c.

$$p(a,b\,|\,c) = \frac{p(a,b,c)}{p(c)} = \frac{p(a)\,p(c\,|\,a)\,p(b\,|\,c)}{p(c)} = p(a\,|\,c)\,p(b\,|\,c)$$

where we used Bayes' theorem $p(c\,|\,a) = p(a\,|\,c)\,p(c)/p(a)$.

Therefore $a \perp\!\!\!\perp b\,|\,c$.

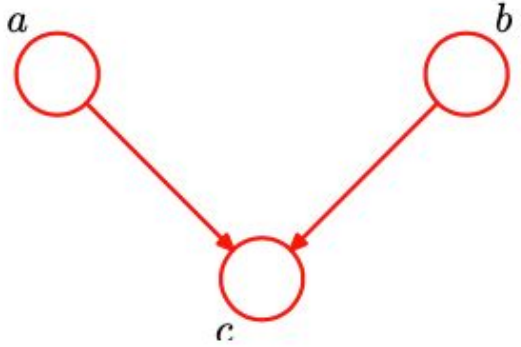
Head-Head (HH) pattern -- a bit more subtle

$$p(a,b,c) = p(a)\,p(b)\,p(c\,|\,a,b)$$

Marginalise over c to test for independence.

$$p(a,b) = \sum_c p(a)\,p(b)\,p(c\,|\,a,b) = p(a)\,p(b)\sum_c p(c\,|\,a,b)$$
$$= p(a)\,p(b)$$

a and b are independent if NO variable is observed:
$a \perp\!\!\!\perp b\,|\,\emptyset$.

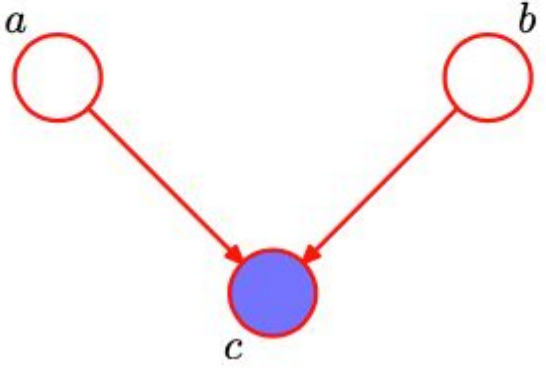


Head-Head (HH) pattern -- 2 of 4

- Now condition on c.

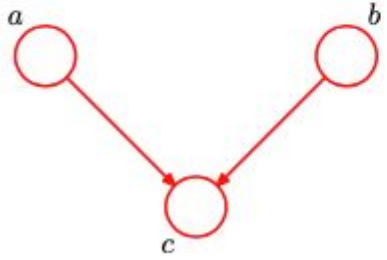
$$p(a,b\,|\,c) = \frac{p(a,b,c)}{p(c)} = \frac{p(a)\,p(b)\,p(c\,|\,a,b)}{p(c)} \neq p(a\,|\,c)\,p(b\,|\,c).$$

- Does not hold : $a \perp\!\!\!\perp b\,|\,c$.

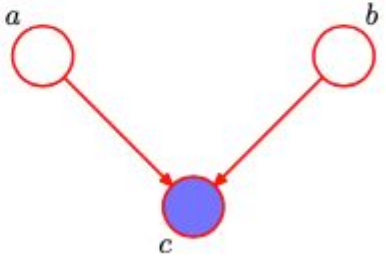


Graphical interpretation

- In both graphical models there is a path from a to b.
- The node c is called head-to-head (HH) with respect to this path because the node c is connected to the heads of the arrows in the path.
- The presence of the HH-node c in the path left makes a independent of b (and b independent of a). The unobserved c blocks the path from a to b.



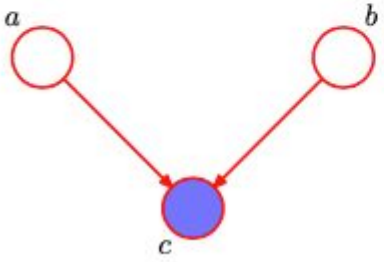
$a \perp\!\!\!\perp b \mid \emptyset$



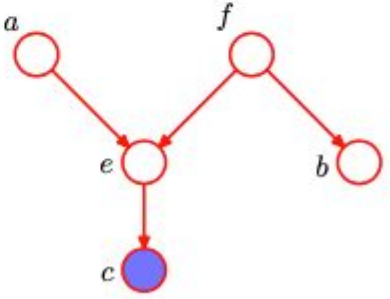
Not $a \perp\!\!\!\perp b \mid c$

Graphical interpretation

- Conditioning on c unblocks the path from a to b, and renders a conditionally dependent on b given c.
- Some more terminology: Node y is a descendant of node x if there is a path from x to y in which each step follows the directions of the arrows.
- A HH-path will become unblocked if either the node, or any of its descendants, is observed.



Not $a \perp\!\!\!\perp b \mid c$



Not $a \perp\!\!\!\perp f \mid c$

HH pattern and “explaining away”

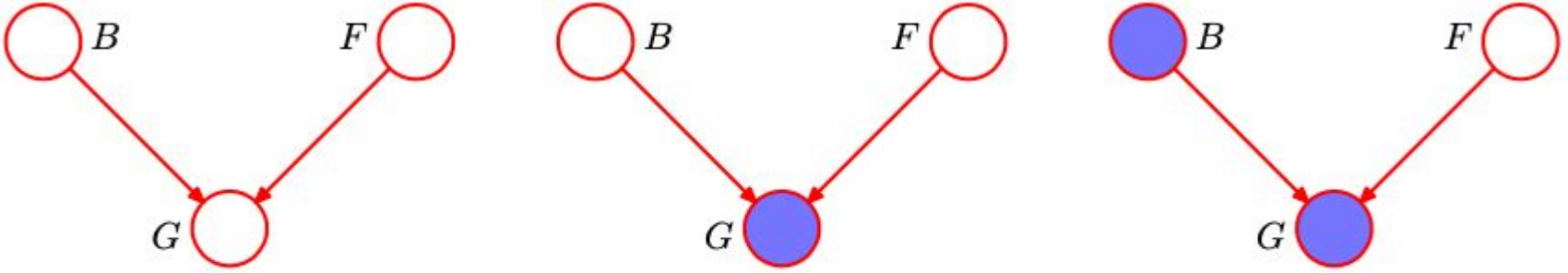


Figure 8.21 An example of a 3-node graph used to illustrate the phenomenon of ‘explaining away’. The three nodes represent the state of the battery (B), the state of the fuel tank (F) and the reading on the electric fuel gauge (G). See the text for details.

Conditional independence for general graphs

- Conditional independence and has been established for
  - all configurations of three variables: (HH, HT, TT) × (observed, unobserved)
  - the subtle case of HH junction with observed descendent.
- We can generalise these results to arbitrary Bayesian Networks.
- Roughly: the graph connectivity implies conditional independence for those sets of nodes which are **directionally separated**.

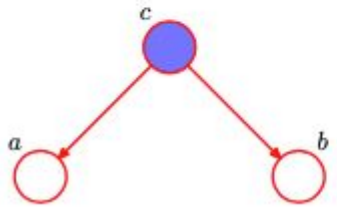
➔ known as “D-separation”

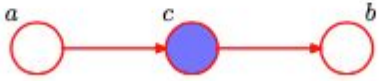
D-separation: blocked paths

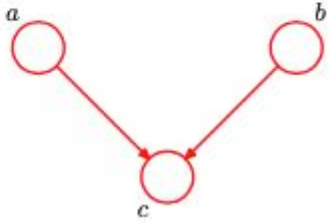
Definition (Blocked Path)

A blocked path is a path which contains

- an observed TT- or HT-node, or
- an unobserved HH-node whose descendants are all unobserved.





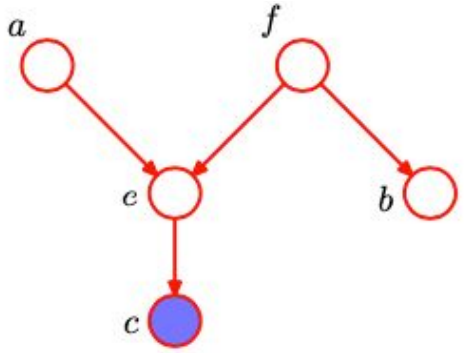


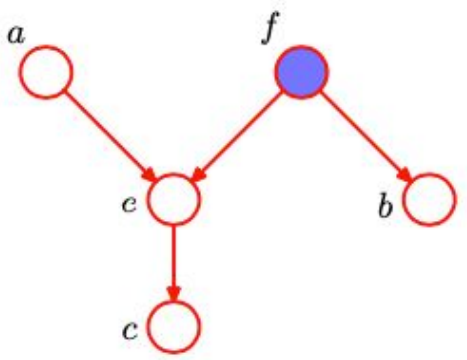
D-separation

- Consider a general directed graph in which A, B, and C are arbitrary non-intersecting sets of nodes. (There may be other nodes in the graph which are not contained in the union of A, B, and C.)
- Consider all possible paths from any node in A to any node in B.
- Any such path is blocked, if it includes a node such that either
  - the node is HT or TT, and the node is in set C, or
  - the node is HH, and neither the node, nor any of the descendants, is in set C.
- If all paths are blocked, then A is $d-$ separated from B by C, and the joint distribution over all the variables in the graph will satisfy $A \perp\!\!\!\perp B \mid C$.

D-separation: example

- The path from a to b is not blocked by f because f is a TT-node and unobserved.
- The path from a to b is not blocked by e because e is a HH-node, and although unobserved itself, one of its descendants (node c) is observed.





- The path from a to b is blocked by f because f is a TT-node and observed. Therefore, $a \perp\!\!\!\perp b \,|\, f$.
- Furthermore, the path from a to b is also blocked by e because e is a HH-node, and neither it nor its descendants are observed. Therefore $a \perp\!\!\!\perp b \,|\, f$.

Theorem (Factorisation ⇒ Conditional Independence)

If a probability distribution factorises according to a directed acyclic graph, and if A, B and C are disjoint subsets of nodes such that A is d-separated from B by C in the graph, then the distribution satisfies $A \perp\!\!\!\perp B \mid C$.

Theorem (Conditional Independence ⇒ Factorisation)

If a probability distribution satisfies the conditional independence statements implied by d-separation over a particular directed graph, then it also factorises according to the graph.

Why is Conditional Independence ⇔ Factorisation relevant?

- Conditional Independence statements are usually what a domain expert knows about the problem at hand.
- Needed is a model $p(\mathbf{x})$ for computation.
- The Conditional Independence ⇒ Factorisation provides $p(x)$ from Conditional Independence statements.
- One can build a global model for computation from local conditional independence statements.

Graphical model as filters for probability distributions

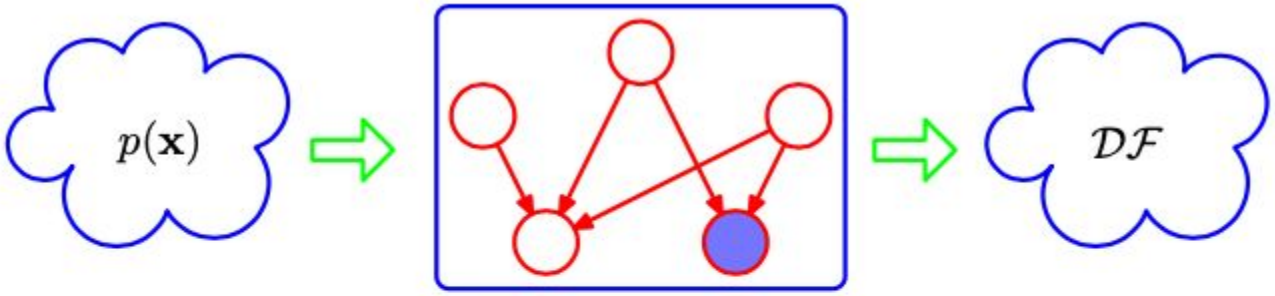


Figure 8.25 We can view a graphical model (in this case a directed graph) as a filter in which a probability distribution $p(\mathbf{x})$ is allowed through the filter if, and only if, it satisfies the directed factorization property (8.5). The set of all possible probability distributions $p(\mathbf{x})$ that pass through the filter is denoted $\mathcal{DF}$. We can alternatively use the graph to filter distributions according to whether they respect all of the conditional independencies implied by the d-separation properties of the graph. The d-separation theorem says that it is the same set of distributions $\mathcal{DF}$ that will be allowed through this second kind of filter.

Graphical models: Bayes Nets

What? And Why?

What is graphical models / Bayesian network

Plate notation

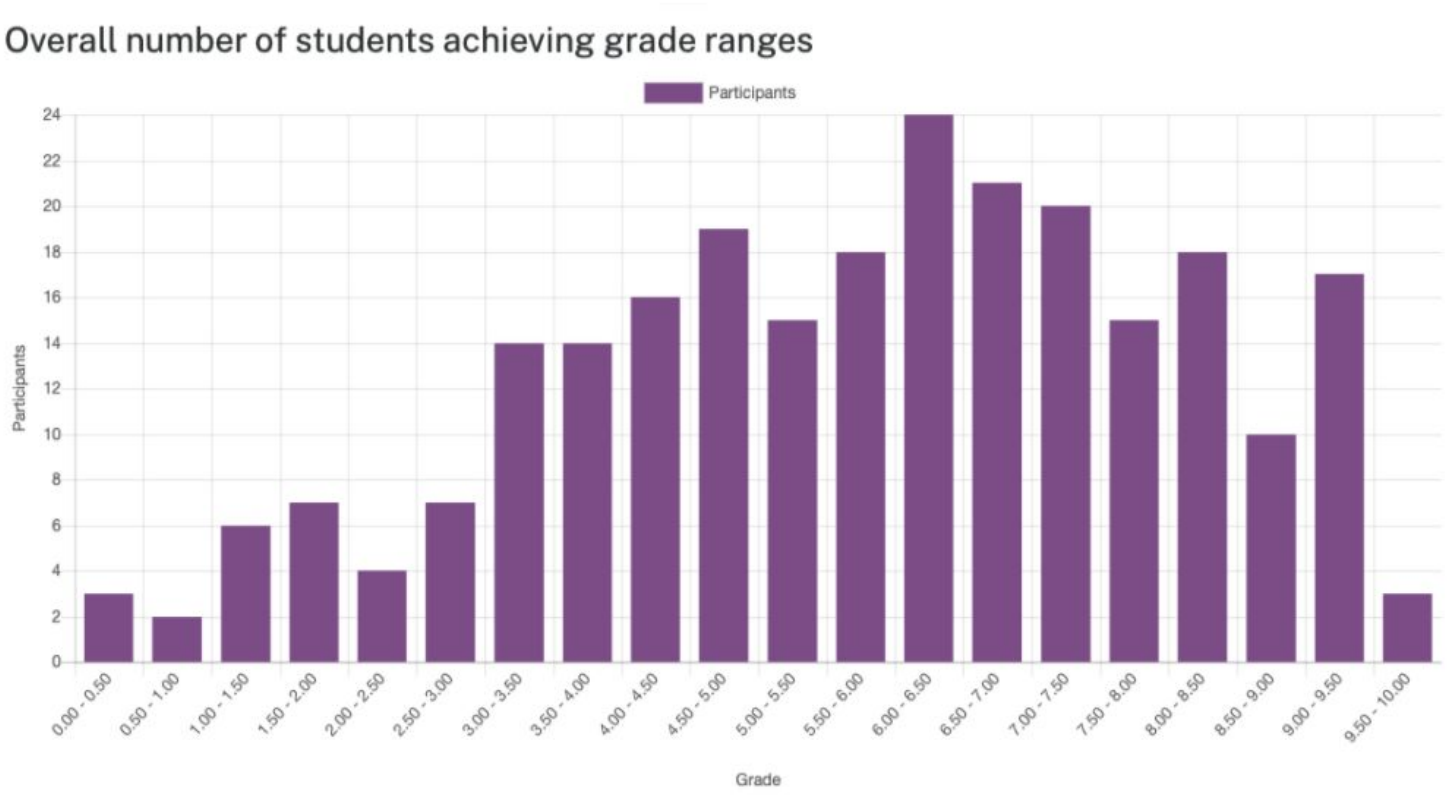
Conditional independence

note @351

quiz 2 stats

253 students completed quiz 2, here are some stats:
average* 57.3; median 60 (out of 100)

the easiest question: Q5 linear regression of gaussian variables 81%; Q8, hoeffding's 71%
the hardest question: Q4 KDE, 23%

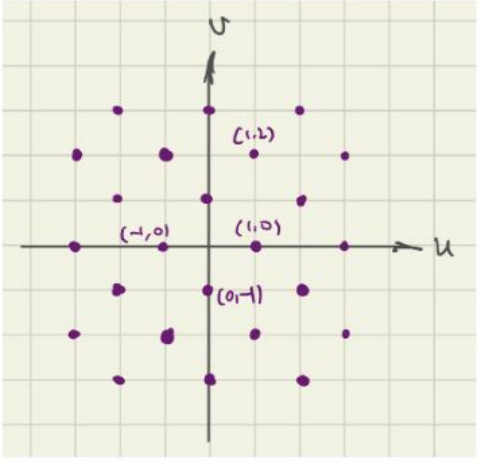


Consider points on a regular grid in the (u, v) plane. The four data points closest to the origin are $(1, 0), (-1, 0), (0, 1), (0, -1)$. For every data point $\mathbf{x}_i = (u_i, v_i)$, there are four other data points at $(u_i \pm 2, v_i)$, and $(u_i, v_i \pm 2)$.

Define the L_1, or manhattan distance between two points as $||\mathbf{x}_i - \mathbf{x}_j||_1 = |u_i - u_j| + |v_i - v_j|$. We perform kernel density estimation using two different kernels $k_1(\mathbf{x}, \mathbf{x}_0)$ and $k_2(\mathbf{x}, \mathbf{x}_0)$, and obtain estimates $p_1(u, v)$ and $p_2(u, v)$, respectively. Those two kernels are defined as following:

Cube kernel: $k_1(\mathbf{x}, \mathbf{x}_0) = \begin{cases} \frac{1}{\sqrt{2}}, & \text{if} \quad ||\mathbf{x} - \mathbf{x}_0||_1 < 1. \\ 0, & \text{otherwise.} \end{cases}$

Pyramid kernel: $k_2(\mathbf{x}, \mathbf{x}_0) = \max\{0, 1 - ||\mathbf{x} - \mathbf{x}_0||_1\}$



Which of the following statements are correct?

Select one or more:

- [ ] A. None of the other options
- [ ] B. $p_2(u, v)$ is continuous everywhere
- [ ] C. The estimates $p_2(u, v)$ is larger or equal to $p_1(u, v)$ everywhere
- [ ] D. Both kernels are computed from data points within a circle in Euclidean space, centred around $\mathbf{x}_0$.
- [ ] E. $p_1(0, 0) > p_2(0, 0)$