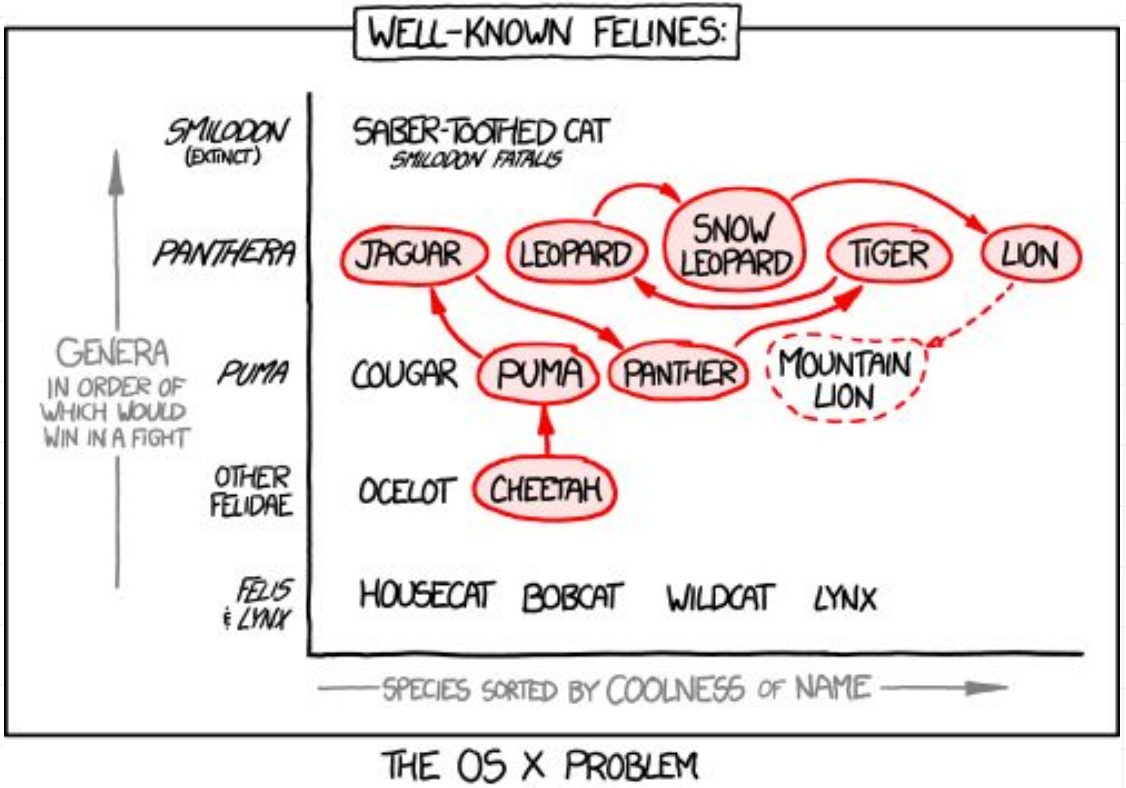
https://xkcd.com/1056/



THE OS X PROBLEM

Announcements:

- Guest lecture next Wed
  PHYS T + Teams
  ML for cyber security
  See you here!

- Two graphical model tutorials:
  week 10 and week 11

- Final exam Fri 3 June
  5:40 - 8:40 pm Canberra
  - On Wattle
  - Self-invigilated, video recording needed
  - Instructions will be available

Markov random fields

Bishop 8.3, 8.4.1-8.4.2

What are MRFs

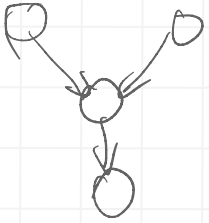
Conditional independence and factorisation

Relation to directed graphs

Inference in graphical models (part 1) - chains and trees

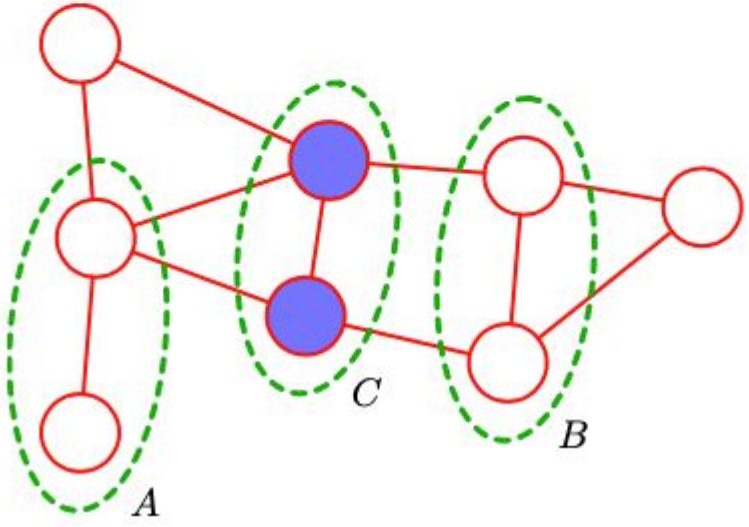
Markov Random fields

- Markov Random Fields (Markov network, undirected graphical model) are defined over a graph with undirected edges.
- MRFs allow for different conditional independence statements than Bayesian networks.
- In a Bayesian network, the definition of a blocked path was subtle for a HH-node because it did include that all descendants were unobservable.
- Is there an alternative graphical semantics for probability distributions such that conditional independence is determined by simple graph separation?
- Yes, removing the direction from the edges removes the asymmetry between parent and child nodes and subsequently the subtleties associated with the HH-node.



Definition: Undirected graph separation

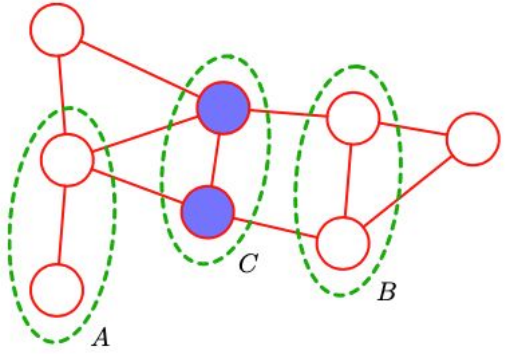
In an undirected graph G, having A, B and C disjoint subsets of nodes, if every path from A to B includes at least one node from C, then C is said to separate A from B in G.



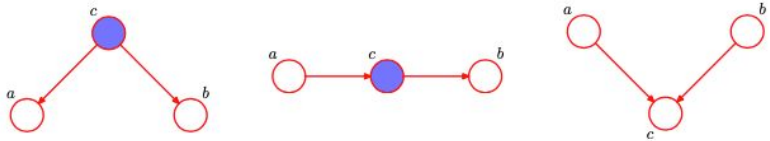
Separation in undirected vs directed graphs

BN, Monday.

In an undirected graph G, having A, B and C disjoint subsets of nodes, if every path from A to B includes at least one node from C, then C is said to separate A from B in G.

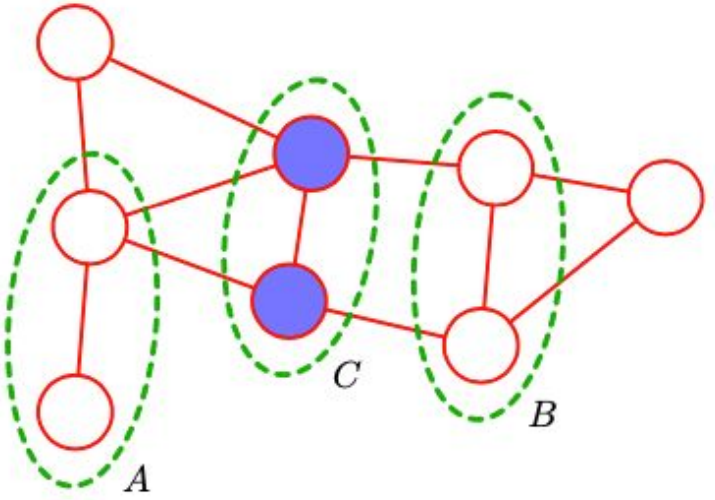


- Consider a general directed graph in which A, B, and C are arbitrary non-intersecting sets of nodes. (There may be other nodes in the graph which are not contained in the union of A, B, and C.)
- Consider all possible paths from any node in A to any node in B.
- Any such path is blocked, if it includes 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 $A \perp\!\!\!\perp B \mid C$.



Conditional independence in MRFs

Figure 8.27 An example of an undirected graph in which every path from any node in set A to any node in set B passes through at least one node in set C. Consequently the conditional independence property $A \perp\!\!\!\perp B \mid C$ holds for any probability distribution described by this graph.



Definition (Markov Random Field)

A Markov Random Field is a set of probability distributions $\{p(\mathbf{x}) \mid p(\mathbf{x}) > 0, \forall \mathbf{x}\}$ such that there exists an undirected graph G with disjoint subsets of nodes A, B and C, in which whenever C separates A from B in G,

$$A \perp\!\!\!\perp B \mid C.$$

- Although we sometimes say "the MRF is such an undirected graph", we mean "the MRF represents the set of all probability distributions whose conditional independency statements are precisely those given by graph separation in the graph".

Factorisation in an MRF

- Assume two nodes x_i and x_j that are not connected by an edge.
- Given all other nodes in the graph, x_i and x_j must be conditionally independent as all paths between x_i and x_j are blocked by observed nodes.

$$p(x_i, x_j|\mathbf{x}_{\setminus\{i,j\}}) = p(x_i|\mathbf{x}_{\setminus\{i,j\}})p(x_j|\mathbf{x}_{\setminus\{i,j\}}) \quad (8.38)$$

where $\mathbf{x}_{\setminus\{i,j\}}$ denotes the set of all variables $\mathbf{x}$ with x_i and x_j removed.

- This is suggestive of the importance of considering sets of connected nodes (cliques). Can we use this for the factorisation of the graph?

$x_i \perp\!\!\!\perp x_j \mid \vec{x}_{\setminus\{i,j\}}$

all nodes except x_i, x_j

Cliques in graphs

A clique is a subset of nodes in a graph such that there exists an edge between all pairs of nodes in the subset. (The nodes in a clique are fully connected.)

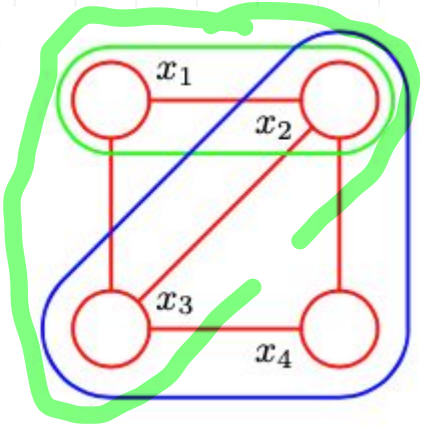
A maximal clique of a graph is a clique which is not a proper subset of another clique. (No other nodes of the graph can be added to a maximal clique without destroying the property of full connectedness.)

Figure 8.29 A four-node undirected graph showing a clique (outlined in green) and a maximal clique (outlined in blue).

$\psi(x_1, x_2)$

$\psi(x_1, x_2, x_3)$

$\psi(x_2, x_3, x_4)$



Factorisation using cliques

x_C : nodes in maximal clique

A probability distribution $p(\mathbf{x})$ factorises with respect to a given undirected graph G if it can be written as

$$p(\mathbf{x}) = \frac{1}{Z} \prod_C \psi_C(\mathbf{x}_C). \qquad (8.39)$$

where $\mathcal{C}$ is the set of maximal cliques of G, and potential functions $\psi_C(\mathbf{x}_C) \geq 0$. The constant $Z = \sum_{\mathbf{x}} p(\mathbf{x})$ ensures the correct normalisation of $p(\mathbf{x})$.

Potential function

$$\psi_C(\mathbf{x}_C) \geqslant 0$$

Normalisation factor / partition function

$$Z = \sum_{\mathbf{x}} \prod_C \psi_C(\mathbf{x}_C) \qquad (8.40)$$

Major limitation of MRFs, computing this for the entire graph is usually hard.

Conditional independence and factorisation

Theorem (Factorisation ⇒ Conditional Independence)

If a probability distribution factorises according to an undirected graph, and if A, B and C are disjoint subsets of nodes such that C separates A from B in the graph, then the distribution satisfies $A \perp\!\!\!\perp B \mid C$.

Theorem (Conditional Independence ⇒ Factorisation (Hammersley-Clifford Theorem)) (1990)

If a strictly positive probability distribution $p(\mathbf{x}) > 0, \forall \mathbf{x}$, satisfies the conditional independence statements implied by graph separation over a particular undirected graph, then it also factorises according to the graph.

Strictly positive potential functions

As the potential functions $\psi_C(\mathbf{x}_C)$ are strictly positive, one can express them as exponential

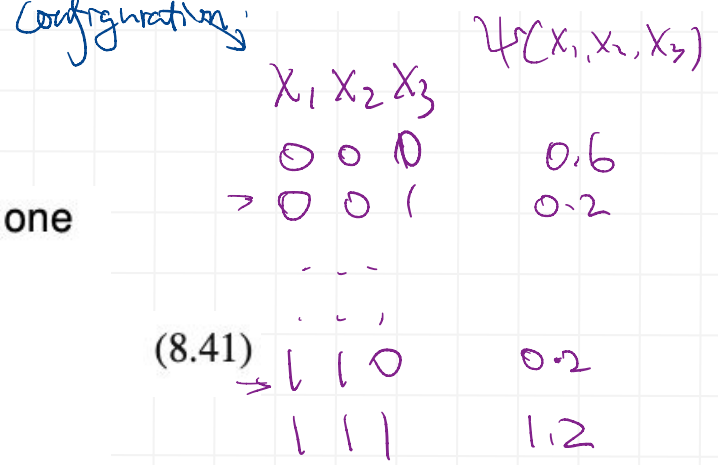
$$\psi_C(\mathbf{x}_C) = \exp\{-E(\mathbf{x}_C)\} \tag{8.41}$$

of an energy function $E(\mathbf{x}_C)$.

The exponential distribution is called the Boltzmann distribution.

The joint distribution is defined as the product of the potentials, and so the total energy is obtained by adding the energies of each of the maximal cliques.

$$p(\mathbf{x}) = \frac{1}{Z}\prod_{C\in\mathcal{C}}\psi_C(\mathbf{x}_C) = \frac{1}{Z}\exp\left\{-\sum_{C\in\mathcal{C}}E(\mathbf{x}_C)\right\}$$



Intuition

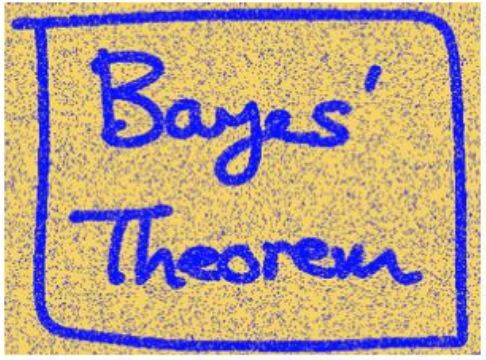
Potential function: expressing which configurations of the local variables are preferred to others.

Global configurations with relatively high probability: those with a good balance in satisfying the (possibly conflicting) influences of the clique potentials.

Example: MRF for image denoising

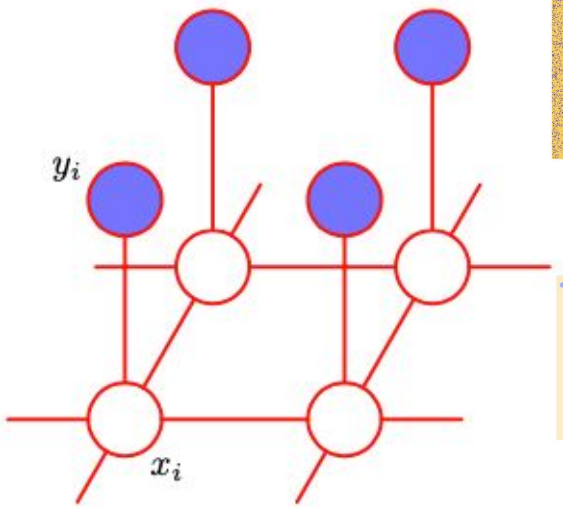
- Given an unknown and noise-free image described by binary pixels $x_i \in \{-1, +1\}$ for $i = 1, \ldots, D$.
- Randomly flip the sign of some pixels with some small probability and denote the pixels of the noisy image by y_i for $i = 1, \ldots, D$.
- Goal : Recover the original noise-free image.

Original image.



After randomly changing 10% of the pixels.

Figure 8.31 An undirected graphical model representing a Markov random field for image de-noising, in which x_i is a binary variable denoting the state of pixel i in the unknown noise-free image, and y_i denotes the corresponding value of pixel i in the observed noisy image.



Bayes' Theorem

Bayes' Theorem

$x_i \in \{-1, 1\}$

A pixel is more likely to be background.

Neighbouring pixels tend to be the same.

A pixel and the corresponding corrupted pixel tend to be the same

$$E(\mathbf{x}, \mathbf{y}) = h \sum_i x_i - \beta \sum_{\{i,j\}} x_i x_j - \eta \sum_i x_i y_i \qquad (8.42)$$

$h > 0$, $\beta > 0$, $\eta > 0$

$$p(\mathbf{x}, \mathbf{y}) = \frac{1}{Z} \exp\{-E(\mathbf{x}, \mathbf{y})\}. \qquad (8.43)$$

1. Fix the elements for $\mathbf{y}$ as we have them observed. (Implicitily defines $p(\mathbf{x} \mid \mathbf{y})$).
2. Initialise $x_i = y_i$ for $i = 1, \ldots, D$.
3. Take one node x_j and evaluate the total energy for both possible states of $x_j = \{-1, +1\}$ keeping all other variables fixed. Set x_j to the state having the lower energy.
4. Repeat for another node until stopping criterion is satisfied.

$\beta = 1.0$, $\eta = 2.1$ and $h = 0$.

Local minimum (ICM).

Global minimum (graph-cut).

Graphical model 2

What are MRFs

Conditional independence and factorisation

Relation to directed graphs

Inference in graphical models (part 1)

Directed vs undirected chain graphs

$$p(\mathbf{x}) = p(x_1)p(x_2|x_1)p(x_3|x_2)\cdots p(x_N|x_{N-1}). \tag{8.44}$$

$$p(\mathbf{x}) = \frac{1}{Z}\psi_{1,2}(x_1,x_2)\psi_{2,3}(x_2,x_3)\cdots\psi_{N-1,N}(x_{N-1},x_N). \tag{8.45}$$

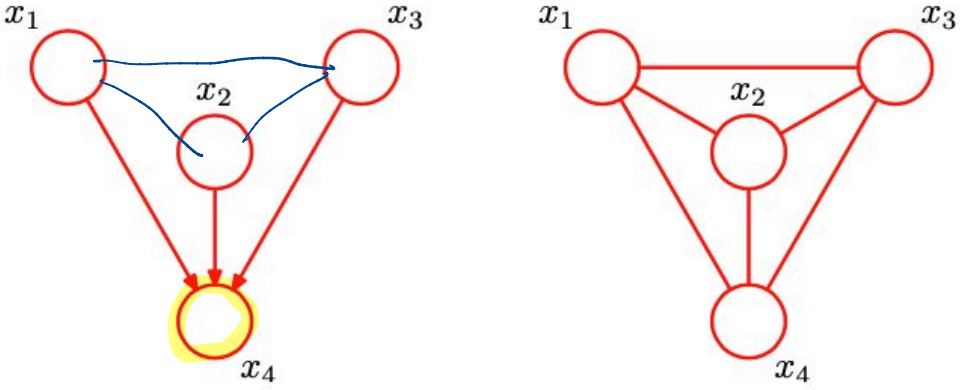
Figure 8.32 (a) Example of a directed graph. (b) The equivalent undirected graph.

$$\begin{aligned}
\psi_{1,2}(x_1,x_2) &= p(x_1)p(x_2|x_1) \\
\psi_{2,3}(x_2,x_3) &= p(x_3|x_2) \\
&\vdots \\
\psi_{N-1,N}(x_{N-1},x_N) &= p(x_N|x_{N-1})
\end{aligned}$$

Directed vs undirected chain graphs have the same factors.
Partition function Z = 1

Moralisation

- For other kind of Bayesian Networks (BNs), create the cliques of a MRF by adding undirected edges between all pairs of parents for each node in the graph.
- This process of 'marrying the parents' is called moralisation, and the result is a moral graph.
- BUT the resulting MRF may represent different conditional independence statements than the original BN.
- Example: The MRF is fully connected, and exhibits NO conditional independence properties, in contrast to the original directed graph.



BN: DAG → no cliques

MRF: all about clique.

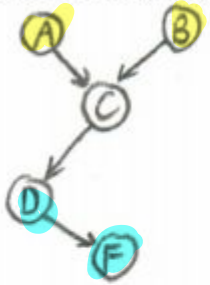
$p(x_1)\,p(x_2)\,p(x_3)$

$p(x_4 \mid x_1, x_2, x_3)$

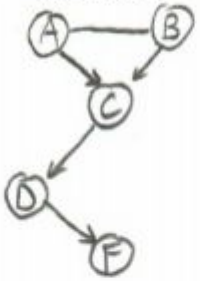
$\psi(x_1, x_2, x_3, x_4)$

1. Are A and B conditionally independent, given D and F?
(Same as "P(A|BDF) =? P(A|DF)" or "P(B|ADF) =? P(B|DF)")

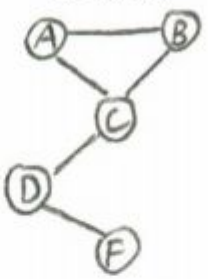
Draw ancestral graph



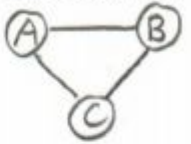
Moralize



Disorient



Delete givens



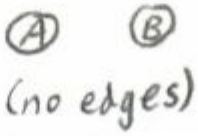
Answer: No, A and B are connected, so they are not required to be conditionally independent given D and F.

2. Are A and B marginally independent? (Same as "P(A|B) =? P(A)" or "P(B|A) =? P(B)")

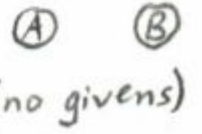
Draw ancestral graph

Moralize

Disorient



Delete givens



Answer: Yes, A and B are not connected, so they are marginally independent.

Example taken from https://web.mit.edu/jmn/www/6.034/d-separation.pdf

Definition (D-map)

A graph is a D-map (dependency map) of a distribution if every conditional independence statement satisfied by the distribution is reflected in the graph.

Definition (I-map)

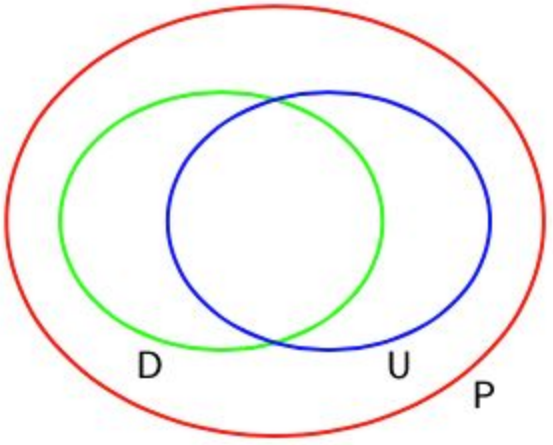
A graph is an I-map (independence map) of a distribution if every conditional independence statement implied by the graph is satisfied in the distribution.

Definition (P-map)

A graph is a P-map (perfect map) of a distribution if it is both a D-map and an I-map for the distribution.

$p(\vec{x})$

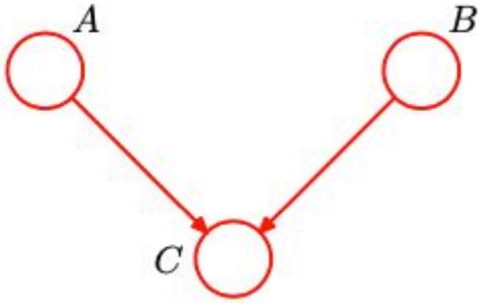
Figure 8.34 Venn diagram illustrating the set of all distributions P over a given set of variables, together with the set of distributions D that can be represented as a perfect map using a directed graph, and the set U that can be represented as a perfect map using an undirected graph.



Undirected : $A \not\perp\!\!\!\perp B \mid C$

Figure 8.35 A directed graph whose conditional independence properties cannot be expressed using an undirected graph over the same three variables.

$A \perp\!\!\!\perp B \mid \phi$



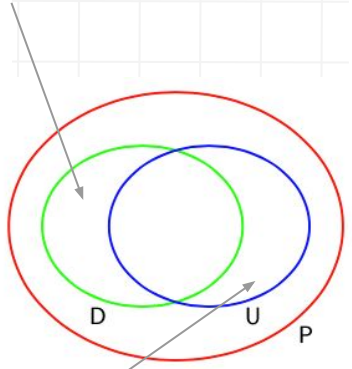
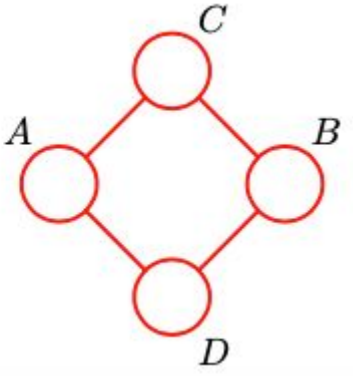


Figure 8.36 An undirected graph whose conditional independence properties cannot be expressed in terms of a directed graph over the same variables.

$A \not\perp\!\!\!\perp B \mid C, D$

$C \perp\!\!\!\perp D \mid A, B$



Bayes Net and MRFs

In both types of Graphical Models

- A relationship between the conditional independence statements satisfied by a distribution and the associated simplified algebraic structure of the distribution is made in term of graphical objects.
- The conditional independence statements are related to concepts of separation between variables in the graph.
- The simplified algebraic structure (factorisation of $p(\mathbf{x})$) is related to 'local pieces' of the graph (child + its parents in BNs, cliques in MRFs).

Differences

- The set of probability distributions that can be represented as MRFs is different from the set that can be represented as BNs.
- Although both MRFs and BNs are expressed as a factorisation of local functions on the graph, the MRF has a normalisation constant $Z = \sum_{\mathbf{x}} \prod_{C \in \mathcal{C}} \psi_C(\mathbf{x}_C)$ that couples all factors, whereas the BN has not.
- The local 'pieces' of the BN are probability distributions themselves, whereas in MRFs they need only be non-negative functions (i.e. they may not have range $[0, 1]$ as probabilities do).

Graphical model 2

What are MRFs

Conditional independence and factorisation

Relation to directed graphs

Inference in graphical models (part 1)

Inference in graphical models

Inference in graphical models: with some of the nodes in a graph are clamped to observed values, compute the posterior distributions of one or more subsets of other nodes.

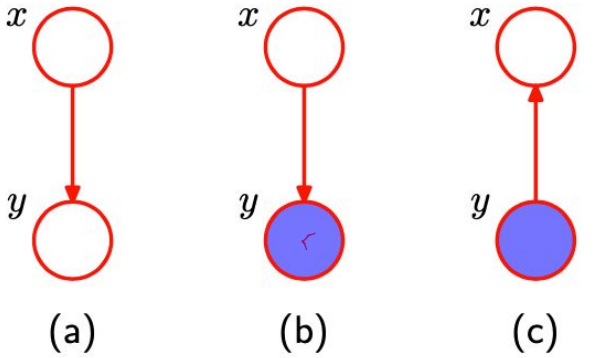
General ideas:

- Exploit graphical structure - efficiency, transparency
- Many algorithms expressed in local "message propagation" around the graph
- Focusing on *exact* inference in this class
  (see Chap 10 for approximate inference algorithms)

Is an essential step in learning (c.f. EM algorithm)

Inference on the simplest graphical model

Figure 8.37 A graphical representation of Bayes' theorem. See the text for details.



$$p(y) = \sum_{x'} p(y|x')p(x') \tag{8.47}$$

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}. \tag{8.48}$$

Inference on a chain graph

Fig 8.32(b)

$$p(\mathbf{x}) = \frac{1}{Z}\psi_{1,2}(x_1, x_2)\psi_{2,3}(x_2, x_3)\cdots\psi_{N-1,N}(x_{N-1}, x_N), \qquad (8.45)$$

also (8.49)

Want to compute $p(x_n)$, naive algorithm: $O(K^{N-1})$

$$p(x_n) = \sum_{x_1}\cdots\sum_{x_{n-1}}\sum_{x_{n+1}}\cdots\sum_{x_N} p(\mathbf{x}). \qquad (8.50)$$

removing x_N

$$\sum_{x_N}\psi_{N-1,N}(x_{N-1}, x_N) \qquad (8.51)$$

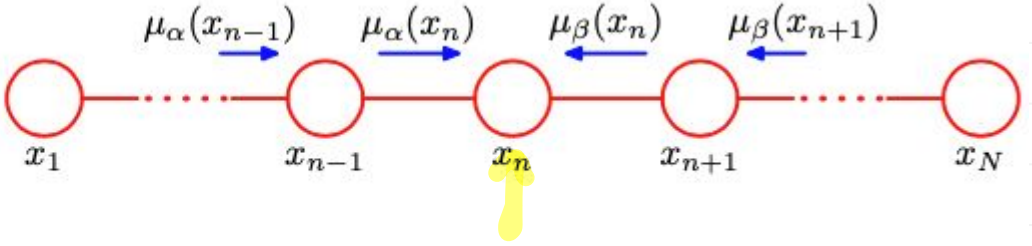
$$p(x_n) = \frac{1}{Z}$$

$$\underbrace{\left[\sum_{x_{n-1}} \psi_{n-1,n}(x_{n-1}, x_n) \cdots \left[\sum_{x_2} \psi_{2,3}(x_2, x_3) \left[\sum_{x_1} \psi_{1,2}(x_1, x_2)\right]\right] \cdots\right]}_{\mu_\alpha(x_n)}$$

$$\underbrace{\left[\sum_{x_{n+1}} \psi_{n,n+1}(x_n, x_{n+1}) \cdots \left[\sum_{x_N} \psi_{N-1,N}(x_{N-1}, x_N)\right] \cdots\right]}_{\mu_\beta(x_n)}. \qquad (8.52)$$

$$p(x_n) = \frac{1}{Z}\mu_\alpha(x_n)\mu_\beta(x_n). \qquad (8.54)$$

Figure 8.38 The marginal distribution $p(x_n)$ for a node x_n along the chain is obtained by multiplying the two messages $\mu_\alpha(x_n)$ and $\mu_\beta(x_n)$, and then normalizing. These messages can themselves be evaluated recursively by passing messages from both ends of the chain towards node x_n.



Recursive evaluation of messages

$$\mu_\alpha(x_2) = \sum_{x_1} \psi_{1,2}(x_1, x_2) \tag{8.56}$$

$$\mu_\alpha(x_3) = \sum_{x_2} \psi_{2,3}(x_2, x_3)\mu_\alpha(x_2)$$

$$\begin{aligned}\mu_\alpha(x_n) &= \sum_{x_{n-1}} \psi_{n-1,n}(x_{n-1}, x_n)\left[\sum_{x_{n-2}} \cdots\right] \\ &= \sum_{x_{n-1}} \psi_{n-1,n}(x_{n-1}, x_n)\mu_\alpha(x_{n-1}).\end{aligned} \tag{8.55}$$

$$\begin{aligned}\mu_\beta(x_n) &= \sum_{x_{n+1}} \psi_{n+1,n}(x_{n+1}, x_n)\left[\sum_{x_{n+2}} \cdots\right] \\ &= \sum_{x_{n+1}} \psi_{n+1,n}(x_{n+1}, x_n)\mu_\beta(x_{n+1}).\end{aligned} \tag{8.57}$$

Other marginals and conditional distributions

brute-force: $O(K^{N-1})$

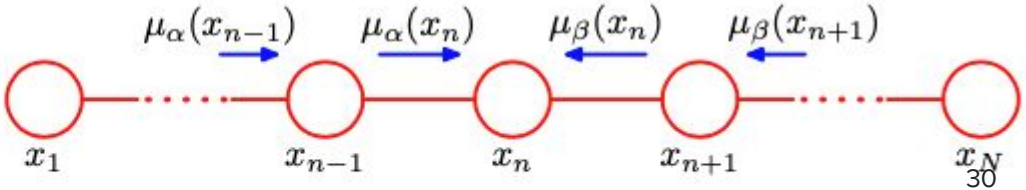
- Computing all marginals
  - Repeat the above N times $O(N^2K^2)$
  - Storing all intermediate messages: $O(N^2K^2)$
- Observed nodes
  - Clamp instead of sum over
- Computing joint probabilities

for each $p(x_n)$ $O(NK^2)$

$p(x_5 | x_3 = 2)$ → K vector.

$p(x_5 | x_3)$ → K×K table

$$p(x_{n-1}, x_n) = \frac{1}{Z}\mu_\alpha(x_{n-1})\psi_{n-1,n}(x_{n-1}, x_n)\mu_\beta(x_n). \qquad (8.58)$$



Tree-shaped graphical models

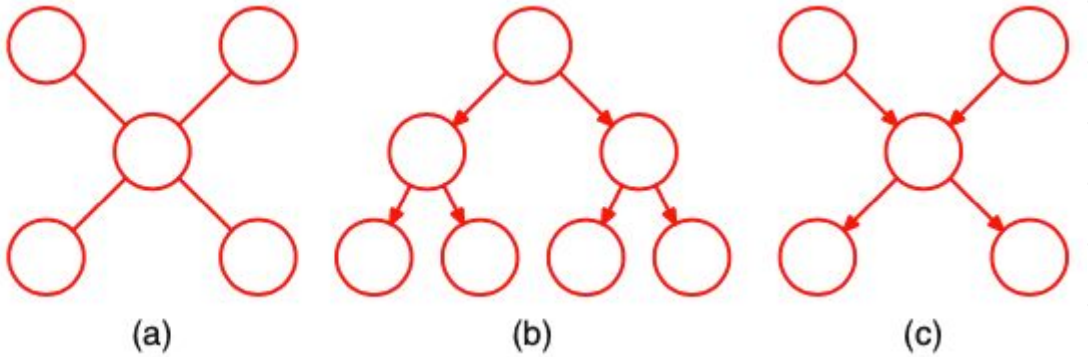
Undirected graph: there is one, and only one, path between any pair of nodes ➔ there is no loops.

Directed graphs: there is a single node, called the root, which has no parents, and all other nodes have one parent.

Moralisation will not add links ➔ convert between directed and undirected tree.

Inference can be done similarly ➔ sum-product algorithm, stay tuned ...

Figure 8.39 Examples of tree-structured graphs, showing (a) an undirected tree, (b) a directed tree, and (c) a directed polytree.



Markov random fields

What are MRFs

Conditional independence and factorisation

Relation to directed graphs

Inference in graphical models (part 1) - chains and trees