Introduction to Graphical Models[1]

Salvador Ruiz Correa

Centro de Investigación en Matemáticas (CIMAT)

[1] These slides are *adapted* from those that accompany the book *Bayesian Reasoning and Machine Learning*. The book and demos can be downloaded from `www.cs.ucl.ac.uk/staff/D.Barber/brml`. We acknowledge David Barber for providing the original slides.

Graphical Models

GMs are graph based representations of various factorization assumptions of distributions. These factorizations are typically equivalent to independence statements amongst (sets of) variables in the distribution.

Belief Network Each factor is a conditional distribution. Generative models, AI, statistics. Corresponds to a DAG.

Markov Network Each factor corresponds to a potential (non negative function). Related to the strength of relationship between variables, but not directly related to dependence. Useful for collective phenomena such as image processing. Corresponds to an undirected graph.

Chain Graph A marriage of BNs and MNs. Contains both directed and undirected links.

Factor Graph A barebones representation of the factorization of a distribution. Often used for efficient computation and deriving message passing algorithms.

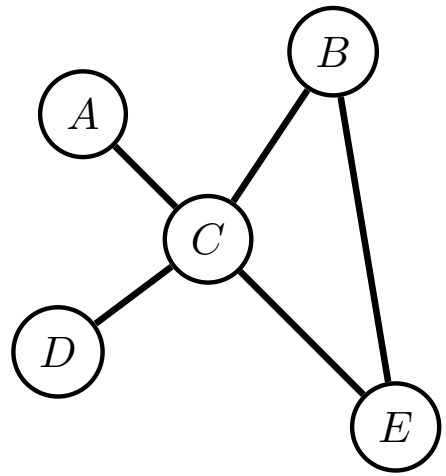
The GM zoo There are many more kinds of GMs, each useful in its own right. We'll touch on some more when we consider inference.

Markov Network

Clique: Fully connected subset of nodes.

Maximal Clique: Clique which is not a subset of a larger clique.

A Markov Network is an undirected graph in which there is a potential (non-negative function) ψ defined on each maximal clique.



The joint distribution is proportional to the product of all clique potentials.

$$p(A, B, C, D, E) = \frac{1}{Z}\psi(A, C)\psi(C, D)\psi(B, C, E)$$

$$Z = \sum_{A,B,C,D,E} \psi(A, C)\psi(C, D)\psi(B, C, E)$$

Z is the normalization factor (partition function). The presence of this constant is one of the major limitations of Markov Networks. Evaluation of this constant is often intractable ($O(K^M)$ for M discrete variables that can take K values).

Markov Network

In general the distribution of a Markov network $\mathcal{H}$ is given by

$$p(x) = \frac{1}{Z} \prod_{x_C \in \mathcal{C}} \psi_C(x_C).$$

where C is a maximal clique of $\mathcal{H}$.

In order to make a formal connection between conditional independence and factorization in Markov networks we need to restrict our attention to potential functions that are strictly positive (i.e. never zero or negative for any choice of the argument variables). Given this restriction we can make a precise relationship between factorization and conditional independence.Because we are restricted to potential functions that are strictly positive it is convenient to express them as exponentials, so that

$$\psi(x_C) = \frac{1}{Z} \exp(-E(x_C)),$$

where $E(x_C)$ is the *energy function* and the exponential representation is called the *Boltzman distribution*:

$$p(x) = \frac{1}{Z} \exp\left(-\sum_{x_C \in \mathcal{C}} E(x_C)\right) = \frac{1}{Z} \exp\left(-E(x)\right).$$

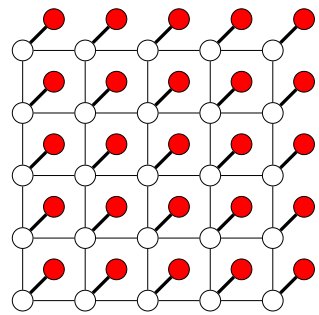
Pairwise Markov Networks

In the special case that the graph contains cliques of only size 2, the distribution is called a *pairwise Markov network*, with potentials defined on each link between two variables.

$$p(x,y) \propto \left[\prod_i \prod_{j \neq i} \psi(x_i, x_j)\right] \left[\prod_i \phi(y_i, x_i)\right]$$

$$= \exp\left(-\sum_i \sum_{j \sim i} E_\psi(x_i, x_j) - \sum_i E_\phi(y_i, x_i)\right)$$

$$= \exp\left(-E(x,y)\right)$$



Computing the most likely x given y. Since $p(x|y) \propto p(x,y)$ we have that

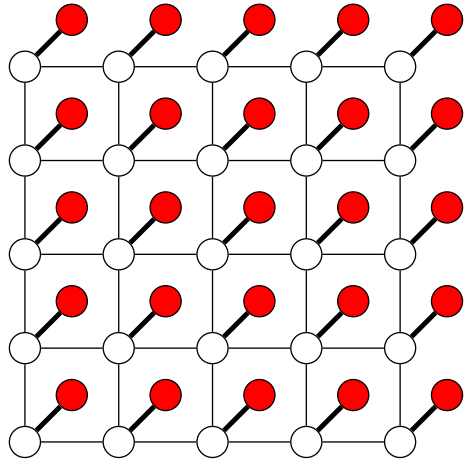
$$x^* = \arg\max_x p(x|y)$$

$$x^* = \arg\max_x p(x,y)$$

$$x^* = \arg\min_x E(x,y) = \arg\min_x \sum_i \sum_{j \sim i} E_\psi(x_i, x_j) + \sum_i E_\phi(y_i, x_i)$$

Example Application of Markov Network – Part I

Problem: We want to recover a binary image from the observation of a corrupted version of it.



$x = \{x_i, i = 1, \ldots, D\}$ $x_i \in \{-1, 1\}$: clean pixel

$y = \{y_i, i = 1, \ldots, D\}$ $y_i \in \{-1, 1\}$: corrupted pixel

$\phi(y_i, x_i) = e^{\gamma x_i y_i}, \quad \gamma > 0$ encourage y_i and x_i to be similar

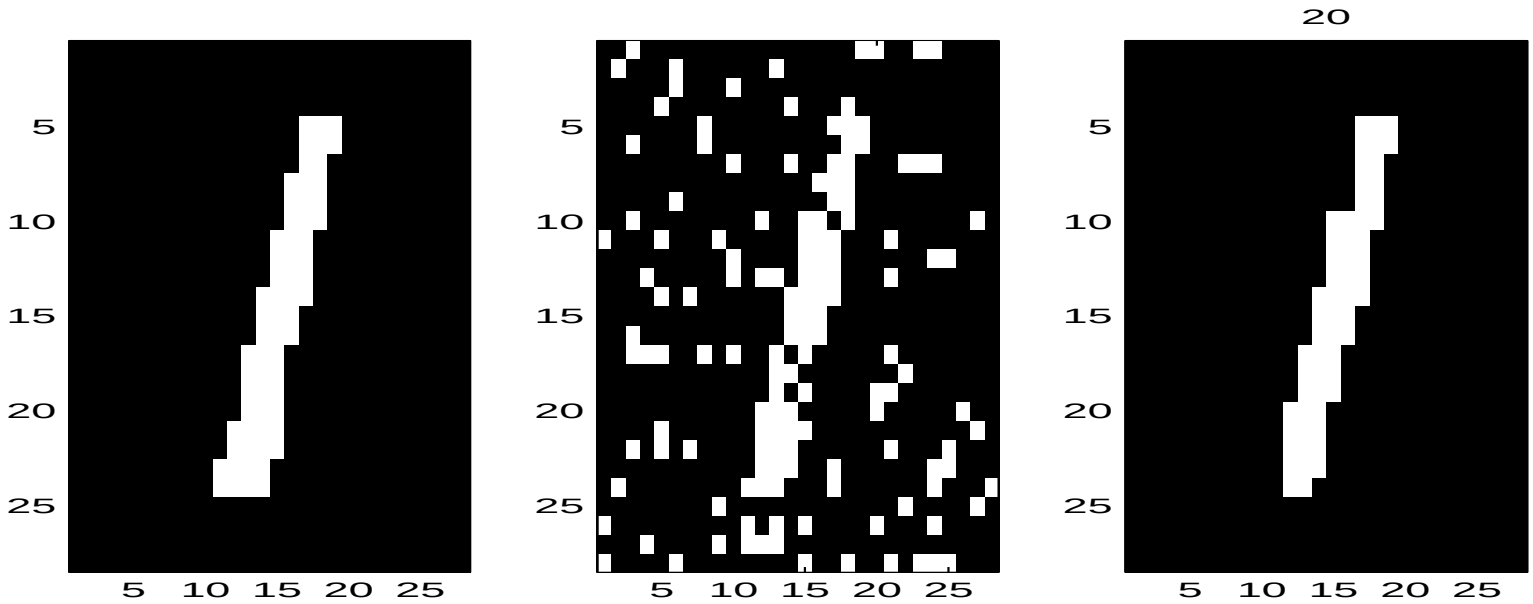
$\psi(x_i, x_j) = e^{\beta x_i x_j}, \quad \beta > 0$ encourage the image to be smooth

$$p(x, y) \propto \left[\prod_{i=1}^{D} \phi(y_i, x_i)\right] \left[\prod_{i \sim j} \psi(x_i, x_j)\right]$$

$$E(x, y) = -\beta \sum_i \sum_{j \sim i} x_i, x_j - \gamma \sum_i y_i x_i$$

Finding the most likely x given y is not easy (since the graph is not singly-connected), but approximate algorithms often work well.

Example Application of Markov Network – Part II



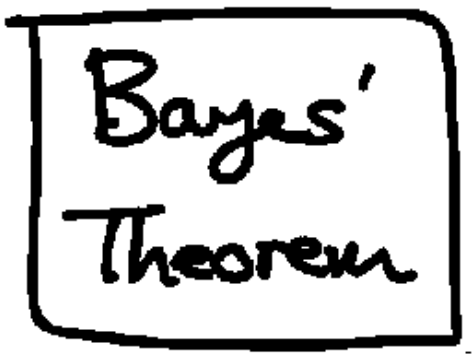
left Original clean image

middle Observed (corrupted) image

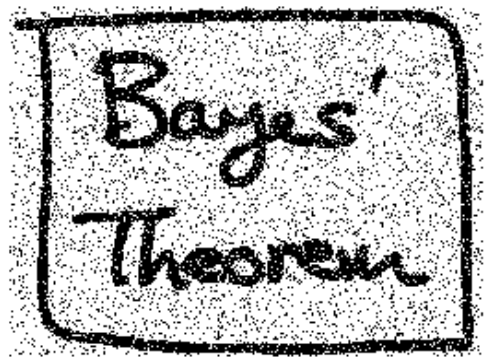
right Most likely clean image $\underset{X}{\operatorname{argmax}}\, p(X|Y)$

Example Application of Markov Network – Part III

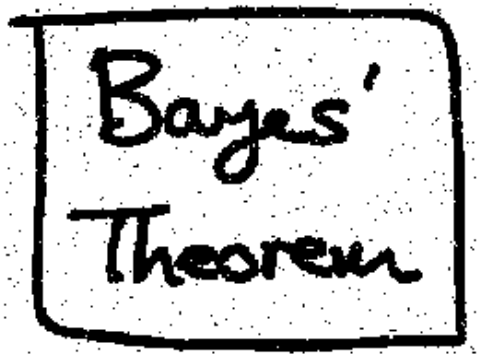
Original



Observed



Most likely image



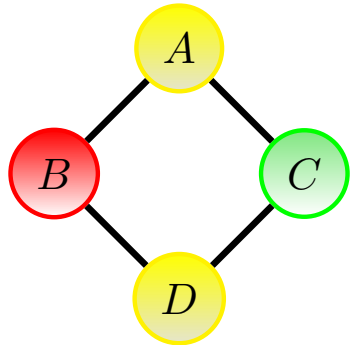
$$p(\text{noise}) = 0.1 \qquad \gamma = 2.1 \qquad \beta = 4.0$$

Example Application of Markov Network – Part IV

Iterated conditional modes. The idea is first to initialize the variables $\{x_i\}$, which we do by simply setting $x_i = y_i$ for all i. Then we take one node x_j at a time and we evaluate the total energy for the two possible states $x_j = +1$ t $x_j = -1$, keeping other node variablees fixed, and set x_j to whichever state has the lower energy. This will either leave the probability unchanged, if x_j is unchanged, or will increase it. Because only one variable is changed, this is a simple local computation that can be performed efficiently. We then repeat the update for another site, and so on, until a suitable stopping criteria us satisfied. The nodes can be updated in a systematic way, for instance by repeatedly raster scanning through the image, or by choosing nodes at random.

If we have a sequence of updates in which every site is visited at least onee, and in which no changes to the variables are made, then by definition the algorithm will have converged to a local maximum of the probability. This need not, however, correspond to the local maximum.

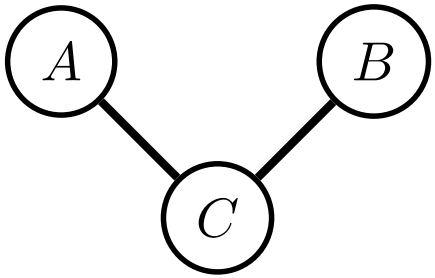
Independence in Markov Networks



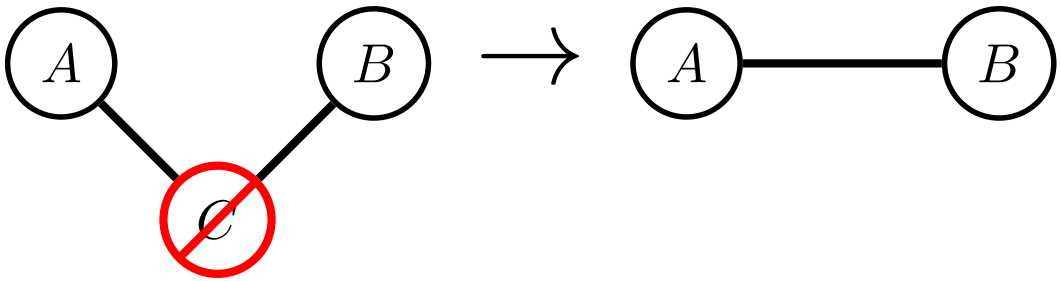
$$B \perp\!\!\!\perp C \mid A, D?$$
$$p(B|A, D, C) = p(B|A, D)?$$

$$\begin{aligned}
p(B|A, D, C) &= \frac{p(A, B, C, D)}{p(A, C, D)} \\
&= \frac{p(A, B, C, D)}{\sum_B p(A, B, C, D)} \\
&= \frac{\psi(A, B)\cancel{\psi(A, C)}\psi(B, D)\cancel{\psi(C, D)}}{\sum_B \psi(A, B)\cancel{\psi(A, C)}\psi(B, D)\cancel{\psi(C, D)}} \\
&= p(B|A, D)
\end{aligned}$$

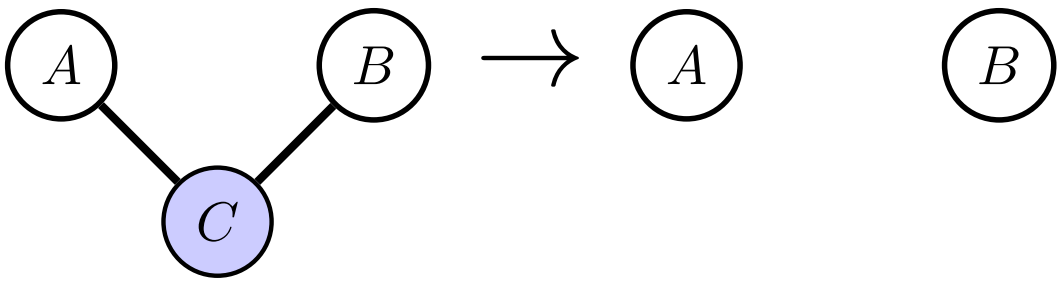
Properties of Markov Networks



$$p(A, B, C) = \phi_{AC}(A, C)\phi_{BC}(B, C)/Z$$

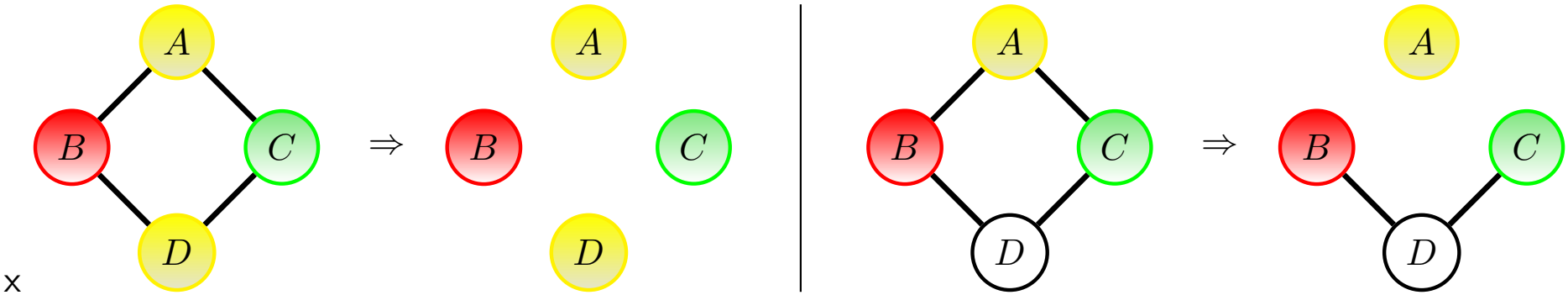


Marginalising over C makes A and B (graphically) dependent. In general $p(A, B) \neq p(A)p(B)$.



Conditioning on C makes A and B independent: $p(A, B|C) = p(A|C)p(B|C)$.

General Rule for Independence in Markov Networks



x

- Remove all links neighbouring the variables in the conditioning set $\mathcal{Z}$.
- If there is no path from any member of $\mathcal{X}$ to any member of $\mathcal{Y}$, then $\mathcal{X}$ and $\mathcal{Y}$ are conditionally independent given $\mathcal{Z}$.

Gibbs distribution

Factor: Let D be a set of random variables. A factor ψ is a function from the image of (D) to $\mathbb{R}$. A factor is not negative if all its entries are non-negative.

A distribution p is a *Gibbs distribution* parametrized by a set of factors $\Psi = \{\psi_1(D_1), \ldots, \psi_K(D_K)\}$ if it is defined as follows:

$$p(x_1, \ldots, x_n) = \frac{1}{Z}\tilde{p}(x_1, \ldots, x_n),$$

where

$$\tilde{p}(x_1, \ldots, x_n) = \psi_1(D_1)\psi_2(D_2)\cdots\psi_K(D_K)$$

is an unnormalized measure and

$$Z = \sum_{x_1,\ldots,x_n} \tilde{p}(x_1, \ldots, x_n)$$

is the partition function.

Factorization over a Markov network. We say that a Gibbs distribution factorizes over a Markov Network $\mathcal{H}$ if each D_k $(k = 1, \ldots, K)$ is a complete subgraph of $\mathcal{H}$ (clique potentials).

Separation

Separation. A subset S separates a subset A from subset B (for disjoint X and Y) in a Markov network $\mathcal{H}$ if every path from any member of X to any member of Y passes through Z. If there is no path from a member X to a member Y then X is separated from Y, which is denoted $\text{sep}_{\mathcal{H}}(X, Y \mid Z)$. If $Z = \emptyset$, then providing that no path exist between X and Y, X and Y are separated.

Separation is monotonic in Z: if $\text{sep}_{\mathcal{H}}(X, Y \mid Z)$ then $\text{sep}_{\mathcal{H}}(X, Y \mid Z')$ for $Z' \supset Z$. Non-monotonic independence relationships cannot be encoded with this definition of separation.

Factorization and I-Map (Soundness of the separation criterion)

Factorization $\Rightarrow$ I-Map:

Let p be a distribution over $\mathcal{X}$, and $\mathcal{H}$ a Markov network structure over $\mathcal{X}$. If p is a Gibbs distribution that factorizes over $\mathcal{H}$, then $\mathcal{H}$ is an I-map for p.

I-Map $\Rightarrow$ Factorization (Hammersley-Clifford theorem):

Let p be a *positive* distribution over $\mathcal{X}$, and $\mathcal{H}$ a Markov network structure over $\mathcal{X}$. If $\mathcal{H}$ is an I-map for p, then p is a Gibbs distribution that factorizes over $\mathcal{H}$.

A distribution is said to be *positive* if for all x in the domain of $\mathcal{X}$ such that $x \neq 0$, $p(x) > 0$.

Markov Network Properties

Global Markov property

$$I(\mathcal{H}) = \{X \perp\!\!\!\perp Y \mid Z : \mathsf{sep}_{\mathcal{H}}(X, Y \mid Z)\}.$$

Local Markov property

$$p(x \mid \mathcal{X} - \{x\}) = p(x \mid \mathsf{Ne}(x)).$$

More specifically,

$$I_l(\mathcal{H}) = \{x \perp\!\!\!\perp \{X - x - \mathsf{Ne}(x)\} \mid \mathsf{Ne}(x) : x \in \mathcal{X}\}.$$

Pairwise Markov property

$$I_p(\mathcal{H}) = \{x \perp\!\!\!\perp y \mid \{X - \{x, y\}\} : \mathsf{edge}(x, y) \notin \mathcal{H}\}.$$

Conditional Independence Properties

- Symmetry: $X \perp\!\!\!\perp Y \mid Z \Rightarrow Y \perp\!\!\!\perp X \mid Z$.
- Decomposition: $X \perp\!\!\!\perp Y, W \mid Z \Rightarrow X \perp\!\!\!\perp Y \mid Z$.
- Weak union: $X \perp\!\!\!\perp Y, W \mid Z \Rightarrow X \perp\!\!\!\perp Y \mid Z, W$.
- Contraction:

$$(X \perp\!\!\!\perp W \mid Z) \ \& \ (X \perp\!\!\!\perp Y \mid Z) \Rightarrow X \perp\!\!\!\perp Y, W \mid Z$$

  .

- Intersection (for positive distributions):

$$(X \perp\!\!\!\perp Y \mid Z, W) \ \& \ (X \perp\!\!\!\perp W \mid Z, Y) \Rightarrow (X \perp\!\!\!\perp Y, W \perp\!\!\!\perp Z)$$

  .

Relationships between Markov Properties – Part I

- For any Markov network $\mathcal{H}$, and any Gibbs distribution p, we have that if p satisfies $I_l(\mathcal{H})$ then it satisfies $I_p(\mathcal{H})$. Proof sketch: Assume that $\text{edge}(x, y) \notin \mathcal{H}$, then

$$\text{sep}_{\mathcal{H}}(x, \mathcal{X} - \{x\} - \text{Ne}(x) \mid \text{Ne}(x)) \;\; \& \;\; \text{sep}_{\mathcal{H}}(y, \mathcal{X} - \{y\} - \text{Ne}(y) \mid \text{Ne}(y))$$

$$\Rightarrow \text{sep}_{\mathcal{H}}(x, y \mid \text{Ne}(x) \cup \text{Ne}(y) \cup R) \quad \text{with } R = \mathcal{X} - \text{Ne}(x) \cup \text{Ne}(y) \cup \{x\} \cup \{y\}$$

$$\Rightarrow \text{sep}_{\mathcal{H}}(x, y \mid \mathcal{X} - \{x\} - \{y\})$$

$$\Rightarrow x \perp\!\!\!\perp y \mid \mathcal{X} - \{x\} - \{y\}$$

$$\Rightarrow p(x, y \mid \mathcal{X} - \{x\} - \{y\}) = p(x \mid \mathcal{X} - \{x\} - \{y\}) p(y \mid \mathcal{X} - \{x\} - \{y\}).$$

Relationships between Markov Properties – Part II

- For any Markov network $\mathcal{H}$, and any distribution p, we have that if p satisfies $I(\mathcal{H})$ it satisfies $I_l(\mathcal{H})$.

  The proof follows directly from the fact that if X and Y are not connected by an edge, then they are necessarily separated by all the remaining nodes of the graph. The converse is true only for positive distributions.

Relationships between Markov Properties – Part III

- Let p be a positive distribution. If p satisfies $I_p(\mathcal{H})$, then p satisfies $I(\mathcal{H})$.

  We need to prove that for all disjoint sets X, Y, Z

  $$\text{sep}_{\mathcal{H}}(X, Y \mid Z) \Rightarrow p \text{ satisfies } \quad X \perp\!\!\!\perp Y \mid Z. \tag{1}$$

  For $|Z| = n - 2$, this equation follows directly from the definition of $I_p(\mathcal{H})$. Using induction, assume that the equation above holds for every Z' with size $|Z'| = k$ and let Z be any set such that $|Z| = k - 1$. We distinguish two cases. In the first case $X \cup Y \cup Z = \mathcal{X}$. In the second case $X \cup Y \cup Z \neq \mathcal{X}$.

Relationships between Markov Properties – Part IV

In the first case, assume, without loss of generality that $|Y| \geq 2$. We have that:

$$Y = Y' \cup A, \quad \text{Y' and A disjoint,}$$

$$X \perp\!\!\!\perp Y' \cup Z,$$

$$X \perp\!\!\!\perp A \cup Z,$$

$$X \perp\!\!\!\perp Y' \mid Z \cup A, \quad \text{as separation is monotonic,}$$

$$X \perp\!\!\!\perp A \mid Z \cup Y', \quad \text{as separation is monotonic.}$$

The separator sets $Z \cup A$ and $Z \cup Y'$ are at least of size $k+1$ and therefore satisfy equation (1). Because p is positive, we can apply the intersection properly $((X \perp\!\!\!\perp Y \mid Z, W) \ \& \ (X \perp\!\!\!\perp W \mid Z, Y) \Rightarrow (X \perp\!\!\!\perp Y, W \mid Z))$ and conclude that p satisfies $X \perp\!\!\!\perp Y' \cup A \mid Z$.

Markov Random Field

A MRF is defined by a set of distributions $p(x_, \mid \mathrm{Ne}(x_i)))$ where $i \in \{1, 2, \ldots, n\}$ indexes the distribution and $\mathrm{Ne}(x)$ are the neighbors of x_i, namely, that the subset of variables x_1,...,x_n that the distribution of variable x_i depends on. The term Markov indicates that this is a proper subset of the variables. A distribution is an MRF with respect to a Markov network $\mathcal{H}$ if

$$p(x_i \mid x_{-i}) = p(x_i \mid \mathrm{Ne}(x_i)),$$

where $\mathrm{Ne}(x_i)$ are the neighboring variables according to $\mathcal{H}$. The notation x_{-i} is a short and for all the set of variables in $\mathcal{X}$ excluding the variable x_i.

The Boltzmann machine

A MN on binary variables $\mathrm{dom}(x_i) = \{0, 1\}$ of the form

$$p(\mathbf{x}|\mathbf{w}, b) = \frac{1}{Z(\mathbf{w}, b)} e^{\sum_{i<j} w_{ij} x_i x_j + \sum_i b_i x_i}$$

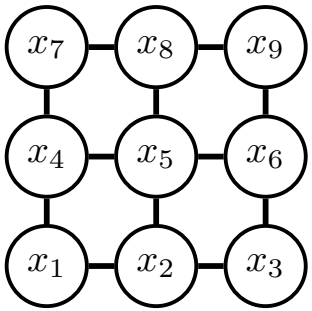
where the interactions w_{ij} are the 'weights' and the b_i the biases.

- This model has been studied in the machine learning community as a basic model of distributed memory and computation. The $x_i = 1$ represents a neuron 'firing', and $x_i = 0$ not firing. The matrix $\mathbf{w}$ describes which neurons are connected to each other. The conditional

$$p(x_i = 1|x_{\setminus i}) = \sigma\left(b_i + \sum_{j \neq i} w_{ij} x_j\right), \qquad \sigma(x) = e^x/(1 + e^x).$$

- The graphical model of the BM is an undirected graph with a link between nodes i and j for $w_{ij} \neq 0$. For all but specially constrained $\mathbf{w}$ inference will be typically intractable.
- Given a set of data $\mathbf{x}^1, \ldots, \mathbf{x}^n$, one can set the parameters $\mathbf{w}, b$ by maximum likelihood (though this is computationally difficult).

The Ising model



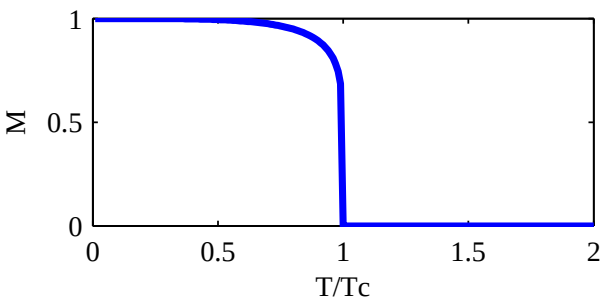
$x_i \in \{+1, -1\}$:

$$p(x_1, \ldots, x_9) = \frac{1}{Z} \prod_{i \sim j} \phi_{ij}(x_i, x_j)$$

$$\phi_{ij}(x_i, x_j) = e^{-\frac{1}{2T}(x_i - x_j)^2}$$

$i \sim j$ denotes the set of indices where i and j are neighbours in the graph. The potential encourages neighbours to be in the same state.

Spontaneous global behaviour

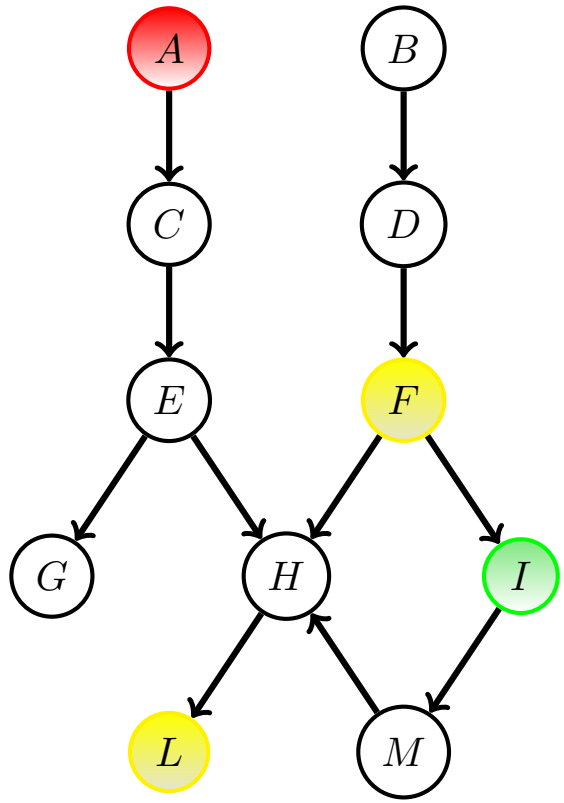


$M = |\sum_{i=1}^{N} x_i|/N$. As the temperature T decreases towards the critical temperature T_c a phase transition occurs in which a large fraction of the variables become aligned in the same state. Even though we only 'softly' encourage neighbours to be in the same state, for a low but finite T, the variables are all in the same state. Paradigm for 'emergent behaviour'.

Alternative Rule for Independence in Belief Networks

$\mathcal{X} \perp\!\!\!\perp \mathcal{Y} | \mathcal{Z}$?

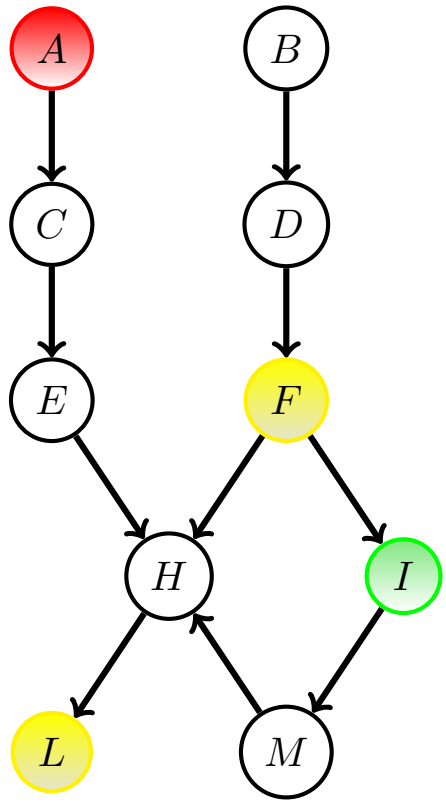
- Ancestral Graph: Remove any node which is neither in $\mathcal{X} \cup \mathcal{Y} \cup \mathcal{Z}$ nor an ancestor of a node in this set, together with any edges in or out of such nodes.
- Moralisation: Add a line between any two nodes which have a common child. Remove arrowheads.
- Separation: Remove all links from $\mathcal{Z}$.
- Independence: If there are no paths from any node in $\mathcal{X}$ to one in $\mathcal{Y}$ then $\mathcal{X} \perp\!\!\!\perp \mathcal{Y} \,|\, \mathcal{Z}$.



Alternative Rule for Independence in Belief Networks

$\mathcal{X} \perp\!\!\!\perp \mathcal{Y} | \mathcal{Z}$?

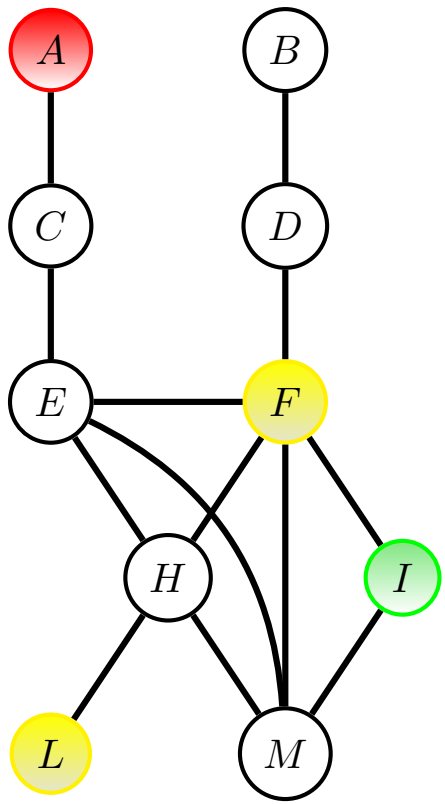
- Ancestral Graph: Remove any node which is neither in $\mathcal{X} \cup \mathcal{Y} \cup \mathcal{Z}$ nor an ancestor of a node in this set, together with any edges in or out of such nodes.
- Moralisation: Add a line between any two nodes which have a common child. Remove arrowheads.
- Separation: Remove all links from $\mathcal{Z}$.
- Independence: If there are no paths from any node in $\mathcal{X}$ to one in $\mathcal{Y}$ then $\mathcal{X} \perp\!\!\!\perp \mathcal{Y} \,|\, \mathcal{Z}$.



Alternative Rule for Independence in Belief Networks

$\mathcal{X} \perp\!\!\!\perp \mathcal{Y} | \mathcal{Z}$?

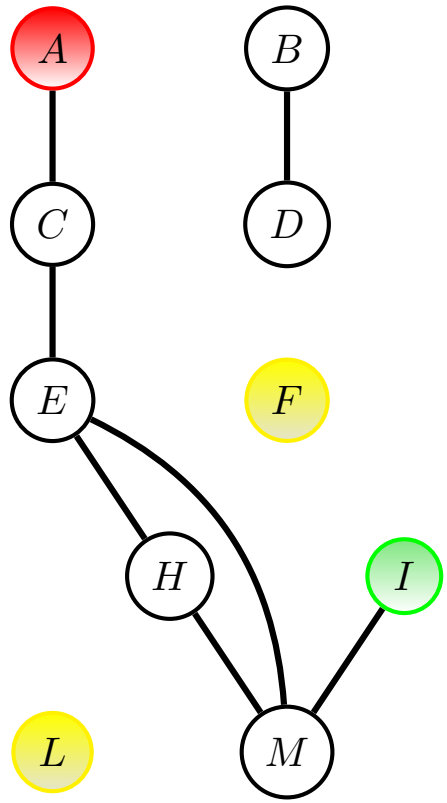
- Ancestral Graph: Remove any node which is neither in $\mathcal{X} \cup \mathcal{Y} \cup \mathcal{Z}$ nor an ancestor of a node in this set, together with any edges in or out of such nodes.
- Moralisation: Add a line between any two nodes which have a common child. Remove arrowheads.
- Separation: Remove all links from $\mathcal{Z}$.
- Independence: If there are no paths from any node in $\mathcal{X}$ to one in $\mathcal{Y}$ then $\mathcal{X} \perp\!\!\!\perp \mathcal{Y} \,|\, \mathcal{Z}$.



Alternative Rule for Independence in Belief Networks

$\mathcal{X} \perp\!\!\!\perp \mathcal{Y} | \mathcal{Z}$?

- Ancestral Graph: Remove any node which is neither in $\mathcal{X} \cup \mathcal{Y} \cup \mathcal{Z}$ nor an ancestor of a node in this set, together with any edges in or out of such nodes.
- Moralisation: Add a line between any two nodes which have a common child. Remove arrowheads.
- Separation: Remove all links from $\mathcal{Z}$.
- Independence: If there are no paths from any node in $\mathcal{X}$ to one in $\mathcal{Y}$ then $\mathcal{X} \perp\!\!\!\perp \mathcal{Y} \,|\, \mathcal{Z}$.

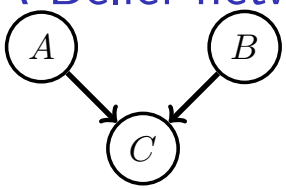


$A \top\!\!\!\top I | F, L$

Expressiveness of Belief and Markov Networks

Cannot represent independence information in certain belief networks with a Markov network.

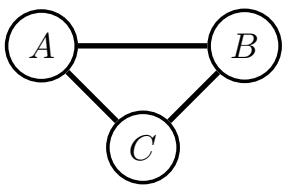
A Belief network



$$A \perp\!\!\!\perp B$$

Markov representation?

Since we have a term $p(C|A,B)$, the MN must have the clique A, B, C:

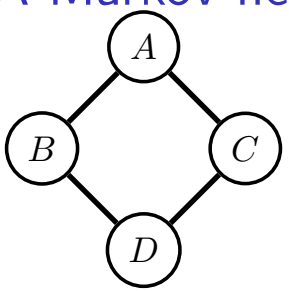


$$A \not\perp\!\!\!\perp B$$

Expressiveness of Belief and Markov Networks

Cannot represent independence information in certain Markov networks with a Belief network.

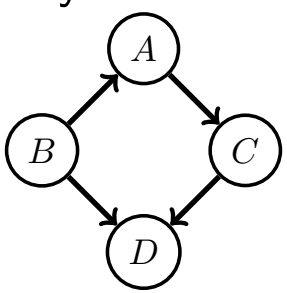
A Markov network



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

Belief Network representation?

Any DAG on A, B, C, D must have a collider.



$$B \top\!\!\!\top C | A, D$$

Representations of distributions

- For a distribution P form list $\mathcal{L}_P$ of all the independence statements.
- For a graph G, form list of all the possible independence statements $\mathcal{L}_G$.

Then we define:

$\mathcal{L}_P \subseteq \mathcal{L}_G$	Dependence Map (D-map)
$\mathcal{L}_P \supseteq \mathcal{L}_G$	Independence Map (I-map)
$\mathcal{L}_P = \mathcal{L}_G$	Perfect Map

In the above we assume the statement l is contained in $\mathcal{L}$ if it is consistent with (can be derived from) the independence statements in $\mathcal{L}$.

Representations of distributions

$$p(t_1, t_2, y_1, y_2) = p(t_1)p(t_2)\sum_h p(y_1|t_1, h)p(y_2|t_2, h)p(h)$$

$$\mathcal{L}_P = \{t_1 \perp\!\!\!\perp (t_2, y_2), \quad t_2 \perp\!\!\!\perp (t_1, y_1)\}$$

Consider the graph of the BN

$$p(y_2|y_1, t_2)p(y_1|t_1)p(t_1)p(t_2)$$

For this we have $\mathcal{L}_G = \{t_2 \perp\!\!\!\perp (t_1, y_1)\}$

- $\mathcal{L}_G \subset \mathcal{L}_P$ so that the BN is an I-MAP for p since every independence statement in the BN is true for the corresponding graph.
- Since $\mathcal{L}_P \not\subseteq \mathcal{L}_G$ the BN is not a D-MAP for p.
- In this case no perfect MAP (a BN or a MN) can represent p.

Representing dependence?

GMs are generally most suited to represented independence. The reason is that local dependence doesn't imply global dependencies. For example

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

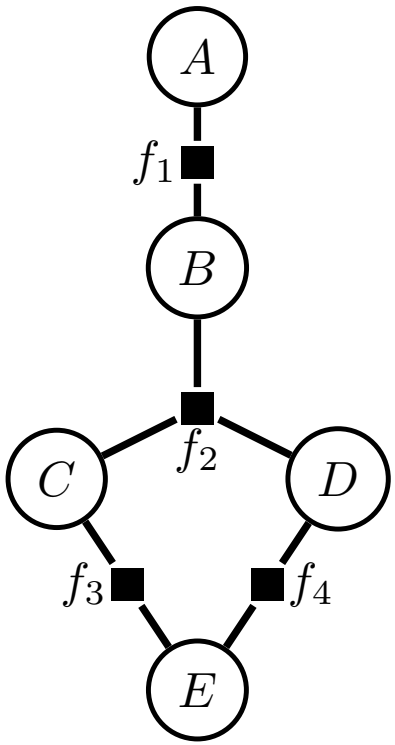
$$p(a) = \begin{pmatrix} 3/5 \\ 2/5 \end{pmatrix}, p(b|a) = \begin{pmatrix} 1/4 & 15/40 \\ 1/12 & 1/8 \\ 2/3 & 1/2 \end{pmatrix}, p(c|b) = \begin{pmatrix} 1/3 & 1/2 & 15/40 \\ 2/3 & 1/2 & 5/8 \end{pmatrix}$$

For these tables, $a \not\perp\!\!\!\perp b$, $b \not\perp\!\!\!\perp c$, but $a \perp\!\!\!\perp c$.

- Local dependence does not guarantee dependence of path-connected variables.
- Graphical independence $\rightarrow$ distribution independence.
- Graphical dependence $\not\rightarrow$ distribution dependence.
- The moral of the story is that graphical models cannot generally enforce distributions to obey the dependencies implied by the graph.

Factor Graphs

A square node represents a factor (non negative function) of its neighbouring variables.

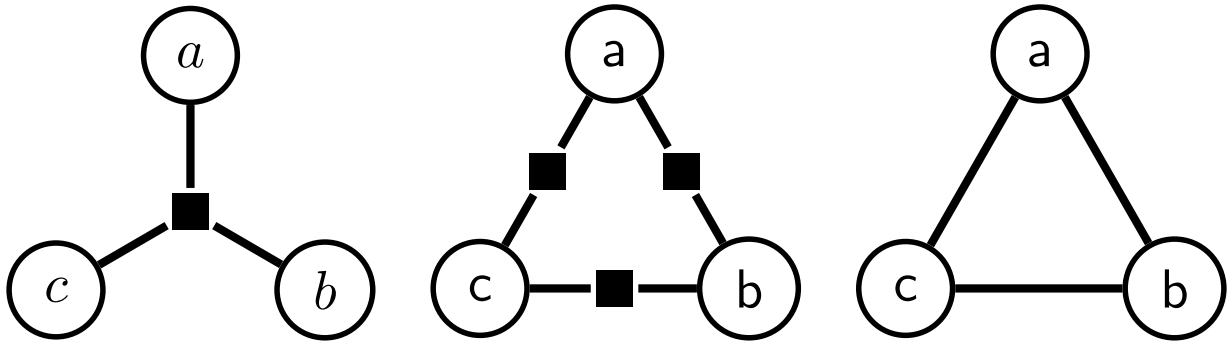


The joint function is the product of all factors:

$$f(A, B, C, D, E) = f_1(A, B) f_2(B, C, D) f_3(C, E) f_4(D, E)$$

Factor graphs are useful for performing efficient computations (not just for probability).

Factor Graphs versus Markov Networks



a $\phi(a, b, c)$

b $\phi(a, b)\phi(b, c)\phi(c, a)$

c $\phi(a, b, c)$

- Both (a) and (b) have the same Markov network (c).
- Whilst (b) contains the same (lack of) independence statements as (a), it expresses more constraints on the form of the potential.