PROBABILISTIC GRAPHICAL MODELS

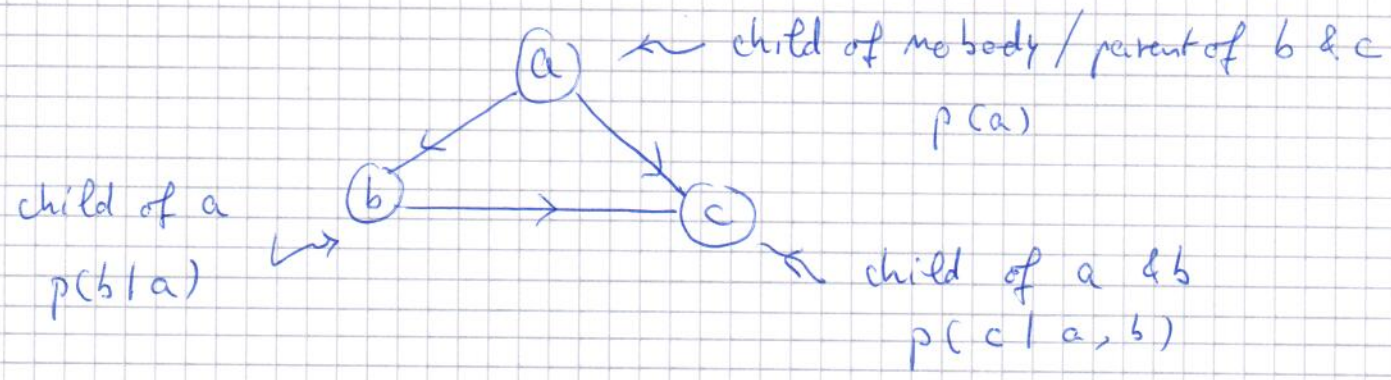
Lecture 1 - BELIEF NETWORKS AND MARKOV RANDOM FIELDS.

I. BELIEF NETWORKS (also called Bayesian Networks).

Let a, b, c three random variables (r.v) with joint distribution $p(a, b, c)$. By Bayes rule we can write

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

The right hand side can be represented graphically as



This representation is called a Directed Acyclic Graph

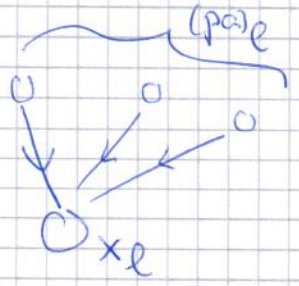
DAG : No cycle is found when you follow the arrows.

The representation is highly non-unique in general. It is always possible to write a joint distr as a DAG in many different ways. The representation becomes useful when the pbch distr has a structure and

some edges are missing corresponding to the fact that some conditional distributions vanish.

Definition: A Belief Network (BN) is a Directed Acyclic Graph (DAG) which defines a joint p.d.f of the form

$$p(x_1, \dots, x_K) = \prod_{\ell=1}^{K} p(x_\ell \mid \underbrace{(pa)_\ell}_{\text{parents of } \ell})$$



Each conditional distribution $p(x_\ell \mid (pa)_\ell)$ is "attached" to node x_ℓ.

Exercise: Let $x_i \in \mathcal{A}$ some discrete alphabet. $\underline{x} = (x_1, \dots x_K) \in \mathcal{A}^K$. Check that for a BN we automatically have a normalization:

$$\sum_{\underline{x} \in \mathcal{A}^K} p(\underline{x}) = 1.$$

hint: Start summing later generations (children) and go upwards in generations.

Examples :

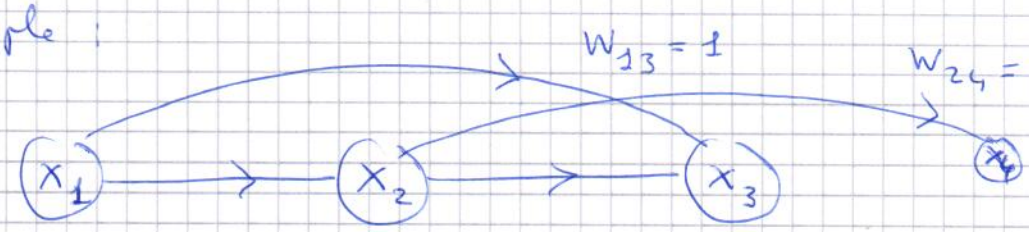
1. Gaussian Belief Network.

$$p(\underline{x}) = \prod_{i=1}^{K} \mathcal{N}\left(x_i \mid \sum_{j \in (pa)_i} w_{ij} x_j + b_i \,;\, v_i^2\right)$$

where

$$\mathcal{N}(x \mid \mu; \sigma^2) = \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}$$

w_{ij} is the adjacency matrix of the DAG : For example :



$w_{12} = 1$ $\quad$ $w_{23} = 1$

In the exercice : compute $\mathbb{E}(x_i)$ & $\mathrm{Cov}(x_i, x_j)$ by recursive marginalisation. You could then learn w_{ij}, b_i, v_i by matching these expressions to empirical moments.

2. Markov Chain

Let $\{X^{(t)}, t = 0 \dots T\}$ a stochastic process with state space $X^{(t)} \in \mathcal{S}$ (the "alphabet").

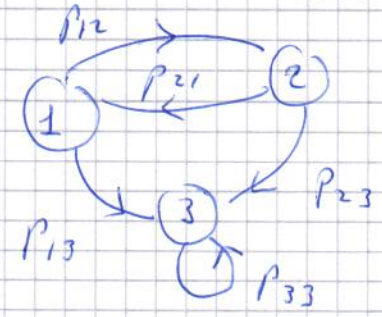
The process is a "Markov Chain" if it satisfies

$$p(X^{(t)} \mid X^{(t-1)} \dots X^{(0)}) = p(X^{(t)} \mid X^{(t-1)}).$$

Furthermore we say it is "homogeneous" if the transition probability satisfies:

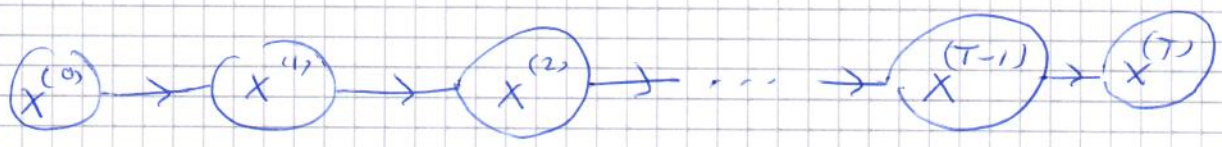
$$p(X^{(t)} = j \mid X^{(t-1)} = i) = p_{i \to j} \quad \text{index of time.}$$

Recall the state graph representation of the Markov chain



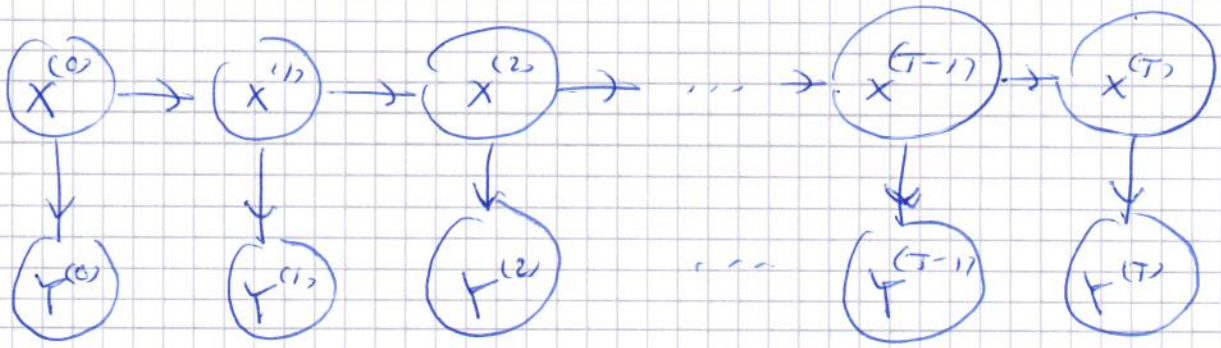
← This is NOT a DAG.

Here we illustrate another representation: The DAG for the joint probability of the MC:



$$p(X^{(0)} \dots X^{(T)}) = p(X^{(0)})\, p(X^{(1)} \mid X^{(0)}) \dots p(X^{(T)} \mid X^{(T-1)}).$$

3. Hidden Markov Model.

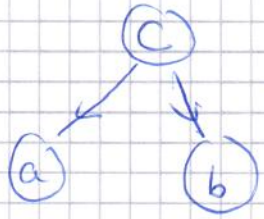


$$\underbrace{p(x^{(0)}, y^{(0)} \dots x^{(T)} y^{(T)})}_{\text{full joint prob}} = p(x^{(0)})\, p(x^{(1)}|x^{(0)}) \dots p(x^{(T)}|x^{(T-1)}) \cdot p(y^{(0)}|x^{(0)})\, p(y^{(1)}|x^{(1)}) \dots p(y^{(T)}|x^{(T)})$$

Here usually this models a situation where a time series $y^{(0)} \dots y^{(T)}$ is visible (observed) and we want to learn the transition probability matrix p_{ij} of a Hidden Markov Chain $(x^{(0)} \dots x^{(T)})$.

II. CONDITIONAL INDEPENCE EXPRESSED BY DAG's.

Example:



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

Claim: $a \perp\!\!\!\perp b \mid c$ "a and b independent given c".

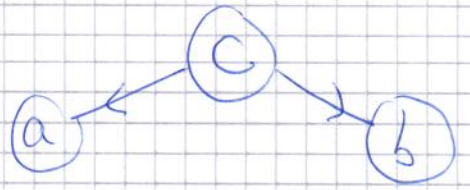
Proof: $p(a,b|c) = \dfrac{p(a,b,c)}{p(c)} = \dfrac{p(a|c)\,p(b|c)\,p(c)}{p(c)}$

$= p(a|c)\,p(b|c)$. ∎

Remark. $p(a,b) \neq p(a)\,p(b)$ so a and b are not independent. They are independent only when conditioned on c.

Three Main situations that can occur in DAG's.

a) Tail to Tail:



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

$a \perp\!\!\!\perp b \mid c$. "a & b indep given c".

b) Head to Tail:

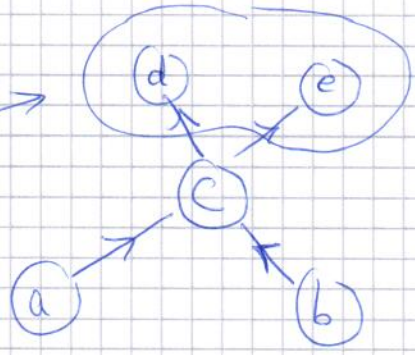
$a \perp\!\!\!\perp b \mid c$ "a & b indep given c".

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

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

c) Head to Head.

descendants of (c).



Here $a \perp\!\!\!\perp b \mid \emptyset$ ← a & b are indep if not conditioned on c neither its descendants

Proof : $p(a,b) = \sum_{c,d,e} p(a,b,c,d,e)$

$= \sum_{c,d,e} p(c|ab)\,p(d|c)\,p(e|c)\,p(a)\cdot p(b)$

$= p(a)\,p(b) \underbrace{\sum_{c} p(c|ab)}_{1} \underbrace{\sum_{d} p(d|c)}_{1} \underbrace{\sum_{e} p(e|c)}_{1}$.

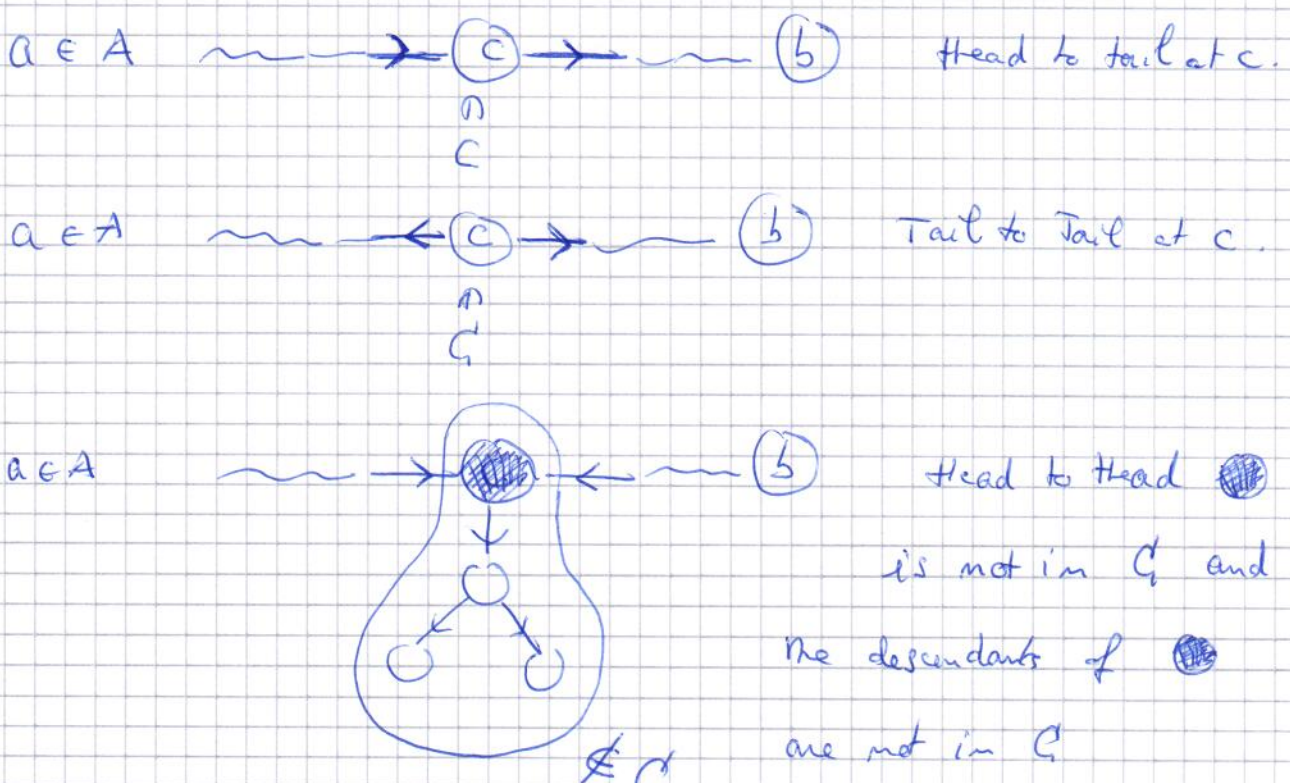
$= p(a)\cdot p(b)$. ∎

General Conditional Indep Principle. Also called sometimes "d-separation principle".

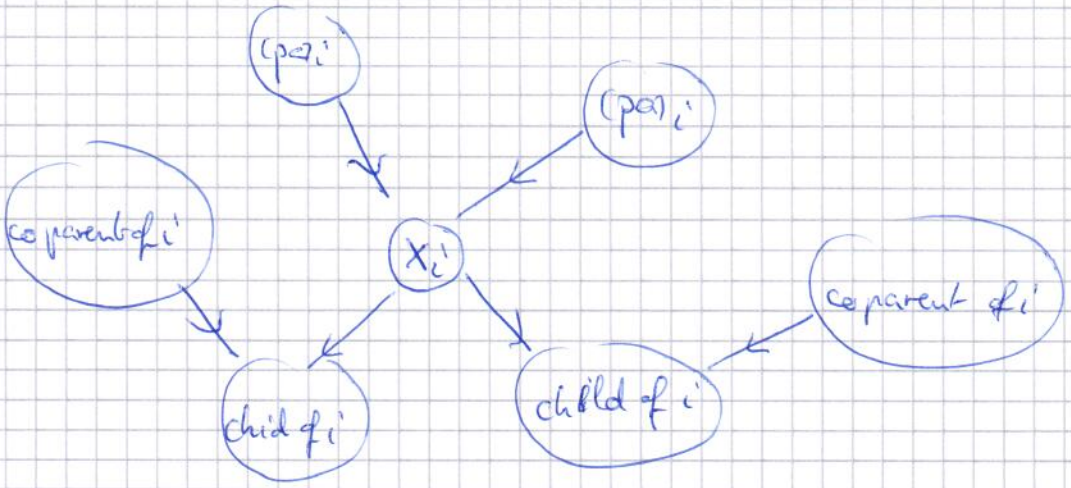
Let A, B, C three disjoint subsets of r.v. If all undirected paths from A to B are "blocked" by G, Then $A \perp\!\!\!\perp B \mid G$

[the converse is not true in general as we can arrange for numerical coincidences].

Definition of "Blocked": A path from (a) to (b) is blocked by G if either of the three happens.



Notion of Markov Blanket of a node in a BN



Markov Blanket of X_i = $MB(i) = \{ (pa)_i, (child)(i), cop(i) \}$.

Lemma : exercise.

$$p(x_i \mid \{x_j\}_{j \neq i}) = p(x_i \mid MB(i))$$

$$p(x_S \mid x_{V \setminus S}) = p(x_S \mid (MB)(S))$$

where S is a set of nodes, $V \setminus S$ the complementary set and $MB(S)$ the Markov Blanket of the set S.

III. MARKOV RANDOM FIELDS (and Gibbs distr).

Markov random fields are undirected graphical models that again allow to express easily independence conditional statements. These are closely related to Gibbs distributions of statistical physics. They are a sort of generalization of Markov Stochastic Process.

Consider $G = (V, E)$ undirected with r.v attached to vertices $\{X_1, X_2, \dots X_N\}$, $N = |V|$ and E a set of edges.

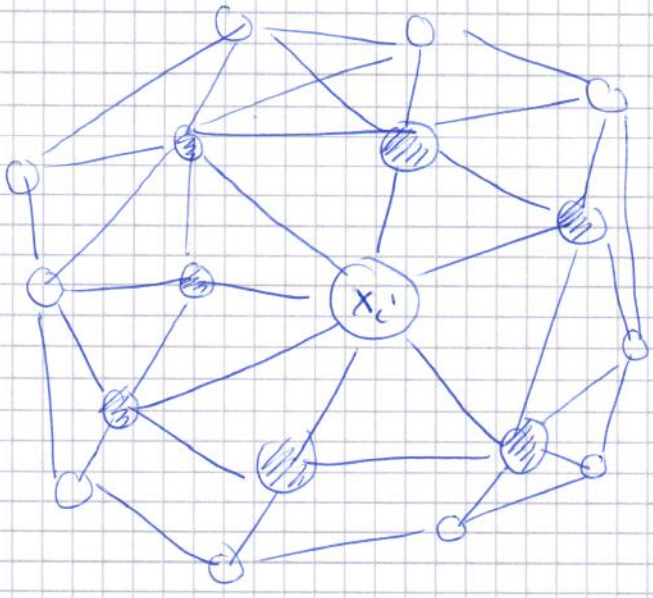
Definition of a MRF:

Given a graph $G = (V, E)$ we say that $p(\underline{x})$ is a MRF if for any set $S \subset V$ with "external boundary" defined by

$$B_S = \{j \in V \mid (i,j) \in E \text{ with } i \in S, j \in V \setminus S\}$$

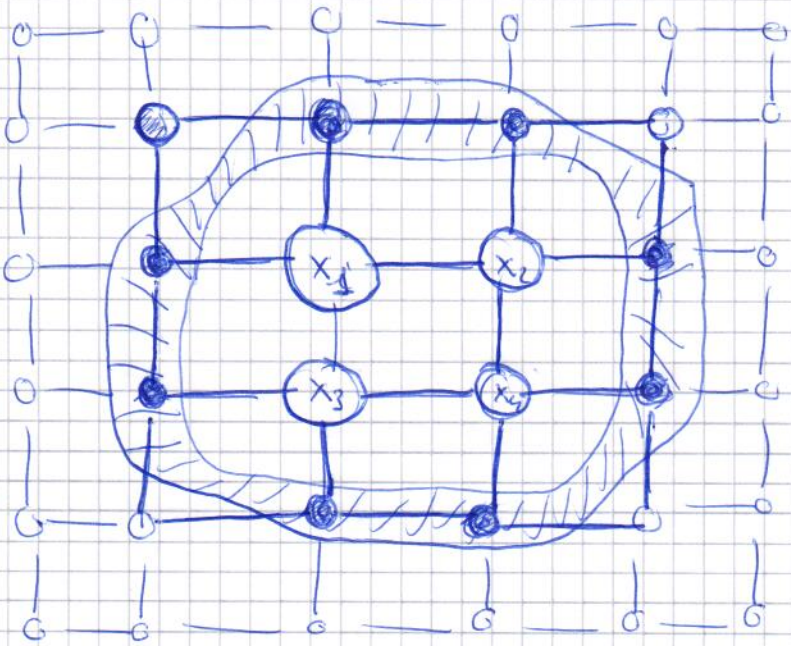
we have

$$p(\{x_i\}_{i \in S} \mid \{x_j\}_{j \in V \setminus S}) = p(\{x_i\}_{i \in S} \mid \{x_j\}_{j \in B_S}).$$



- $G = (V, E)$.
- Here $S = i$.
- $\mathcal{B}_S$ = shaded nodes.
- $\mathcal{B}_S$ is also called the Markov Blanket of S.

$MB(S)$.



- $S = \{x_1, x_2, x_3, x_5\}$
- $MB(S) = \mathcal{B}_S$

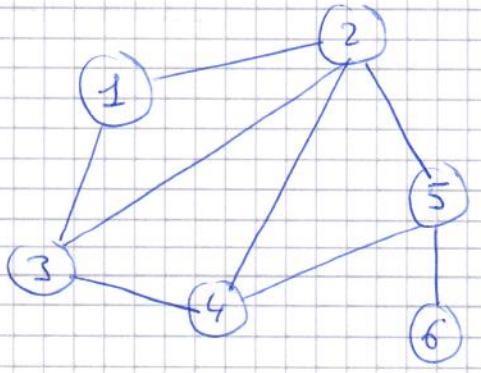
= ● nodes inside

Notation: We adopt the notation $X_S = \{X_i\}_{i \in S}$ for a set of r.v.

Notion of Clique: A clique is a subset $C \subset V$ such that the restriction of $G = (V, E)$ to this subset is a complete graph. All nodes in a clique are pairwise connected.

Maximal Clique: A clique C is maximal if when we add an extra node the set $C \cup \{\text{extra node}\}$ is not a clique.

Example



(1 2), (1 3), (2 3), (2 4), (2 5), (4, 5), (5, 6) are cliques.

(5,6) is also a maximal clique (the other are not maximal).

(1 2 3) maximal clique

(2 4 5) maximal clique

(2 3 4) maximal clique.

Basic Thm for MRF: Hammersley-Clifford ~ 1970's.

Fix $G = (V, E)$, Let $p(\underline{x}) > 0$ (strictly positive otherwise careful!) be a MRF according to the previous definition.

(i) Then $p(\underline{x})$ must be of the form:

$$p(\underline{x}) = \frac{1}{Z} \prod_{\substack{\text{Maximal} \\ \text{cliques} \\ C}} \psi_C(x_C) \qquad (*)$$

where $\psi_C(x_C)$ is a syst of > 0 weights (un-normalized) associated to maximal cliques. Also, Z is called the normalization factor and:

$$Z = \sum_{\underline{x} \in \mathcal{X}^V} \prod_{\substack{\text{Max cliques} \\ C}} \psi_C(x_C).$$

(ii) The converse is also true i.e, $(*) \Rightarrow$ previous conditional indep of MRF.

In other words $(*)$ implies that for any $S \subset V$:

$$p(x_S \mid x_{V \setminus S}) = p(x_S \mid \underline{(MB)(S)}) \text{ or } \underline{B_S}.$$

Remarks:

a). The $\psi_c(x_c)$ are not prob distr. They are positive but not normalized.

b) The normalization factor $\overbrace{Z}$ is also called the partition function.

c) The sum over $\underline{x} \in \mathcal{A}^N$ to compute Z contains an exponential number of terms and is intractable when the number variables grows large.

d) Since $\psi_c(x_c) > 0$ we can always write them as

$$\psi_c(x_c) = \exp\left(-E_c(x_c)\right)$$

where $E_c(x_c) \in \mathbb{R}$ are called "energy functions or "potentials". This representation is used in physics and called the Boltzmann weight repr.

e) $p(\underline{x})$ is called a Gibbs distribution if $\exists$ energy functions $E_S(x_S)$ s.t S is a clique and

$$p(\underline{x}) = \frac{1}{Z} \exp\left\{-\sum_{S \text{ clique}} E_S(x_S)\right\}.$$

usually the sum does not carry necessarily over maximal cliques. For finite lattices MRF $\Leftrightarrow$ Gibbs. For infinite lattices the theory of Gibbs distr is much deeper and goes back to Dobrushin-Lanford-Ruelle (DLR).

f) MRF are a special case of exponential families.

But note that not all exp families are MRF. To see a MRF as an exponential family note that:

$$p(\underline{x}) = \frac{1}{Z} \prod_{\text{max } c} \psi_c(x_c) = \exp\left\{ - \sum_{\text{max } c} \psi_c(x_c) - \log Z \right\}$$

Take $\vartheta_i \to \vartheta_{c, z_c} \equiv - E_c(z_c)$ for $c =$ max clique $z_c \in \mathcal{A}^{|c|}$.

$$T_i(\underline{x}) \to T_{c, z_c}(\underline{x}) \equiv \delta_{x_c z_c}$$

$$i \to (c, z_c) \in \{\text{max cliques}\} \times \mathcal{A}^{|c|}.$$

Then we see that:

$$p(\underline{x}) = h(\underline{x}) \exp\left\{ \sum_i \vartheta_i T_i(\underline{x}) - \psi(\underline{\vartheta}) \right\} \quad (**)$$

with $h(\underline{x}) = 1$ and $\psi(\underline{\vartheta}) = \log Z$.

(**) is the definition of an exponential family with prior $h(\underline{x})$, sufficient statistic $T_i(\underline{x})$ and parameters ϑ_i. The $\exp(-\psi(\vartheta))$ is a normalization factor.

Proof of (ii) in Hammersley-Clifford Thm.

The proof of (i) is non trivial and here we only prove the converse (ii). (See Grimett chap 7 for (i).)

Moreover for simplicity we prove the claim for $S = \{i\}$ a singleton only. So we want to show that the factorization property (*) page 13 implies:

$$p(x_i / \{x_j\}_{j \neq i}) = p(x_i / MB(i))$$

Note that MB(i) = set of max cliques containing node i.

$$l.h.s = p(x_i / \{x_j\}_{j \neq i}) = \frac{p(\underline{x})}{\int dx_i \, p(\underline{x})}$$

$$= \frac{\prod_c \psi_c(x_c)}{\int dx_i \cdot \prod_c \psi_c(x_c)}$$

$$\prod_c \psi_c(x_c) = \left\{ \prod_{c : x_i \in c} \psi_c(x_c) \right\} \left\{ \prod_{c : x_i \notin c} \psi_c(x_c) \right\}$$

$$\int dx_i \prod_c \psi_c(x_c) = \left\{ \int dx_i \prod_{c : x_i \in c} \psi_c(x_c) \right\} \left\{ \prod_{c : x_i \notin c} \psi_c(x_c) \right\}$$

$$\Rightarrow l.h.s = \frac{\prod_{c : x_i \in c} \psi_c(x_c)}{\int dx_i \prod_{c : x_i \in c} \psi_c(x_c)} .$$

$$r.h.s = p(x_i \mid MB(i)) = \frac{p(x_i, MB(i))}{\int dx_i\, p(x_i, MB(i))}$$

$$= \frac{\int dx_{\sim i,\, MB(i)}\, p(\underline{x})}{\int dx_i \int dx_{\sim i,\, MB(i)}\, p(\underline{x})}$$

where $\int dx_{\sim i,\, MB(i)}$ means integral w.r.t to all variables except i, $MB(i)$.

$$\text{Numerator} = \int dx_{\sim i,\, MB(i)}\, p(\underline{x}) = \int dx_{\sim i,\, MB(i)} \underbrace{\prod_{c:\, x_i \in c} \psi_c(x_c)}_{\text{contains } i \text{ \& } MB(i) \text{ only so we can pull it out of integral.}} \prod_{c:\, x_i \notin c} \psi_c(x_c)$$

$$= \prod_{c:\, x_i \in c} \psi_c(x_c) \left\{ \int dx_{\sim i\, MB(i)} \prod_{c:\, x_i \notin c} \psi_c(x_c) \right\}$$

$$\text{Denominator} = \left\{ \int dx_i \prod_{c:\, x_i \in c} \psi_c(x_c) \right\} \left\{ \int dx_{\sim i\, MB(i)} \prod_{c:\, x_i \notin c} \psi_c(x_c) \right\}$$

$$\Rightarrow r.h.s = \frac{\prod_{c:\, x_i \in c} \psi_c(x_c)}{\int dx_i \prod_{c:\, x_i \in c} \psi_c(x_c)}$$

We see that $l.h.s = r.h.s$ ∎ QED ∎

Examples of MRF

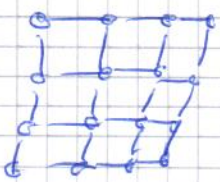
1) Ising Model. (1920's !)

$G = (V, E)$, s_i, $i \in V$ with $s_i \in \{-1, +1\}$ are called "spins" .

$$p(\underline{s}) = \frac{1}{Z} \exp\left\{ - \sum_{(i,j) \in E} J_{ij} s_i s_j - \sum_{i \in V} h_i s_i \right\} .$$

J_{ij} , $h_i \in \mathbb{R}$.

- It can be shown that this is a MRF (or a Gibbs distr.).
- Usually in statistical physics J_{ij}, h_i are given and we want to compute marginals of $p(\underline{s})$. In learning theory we are given samples $\underline{s}^{(1)} \dots \underline{s}^{(N)} \sim p(\underline{s})$ and we want to learn J_{ij}, h_i as well as G. This is also called the "Inverse Ising Problem".
- G = complete graph , $J_{ij} = \frac{J}{|V|}$, $J > 0$ → Curie-Weiss Model
  ∃ One clique ! (max)
- G = Cube $\subset \mathbb{Z}^d$, $J_{ij} = \begin{cases} J & \text{for } |i-j| = 1 \\ 0 & \text{otherwise} \end{cases}$



→ Traditional Ising model on square grid.

→ Application in Image Processing for example magnetism ect...

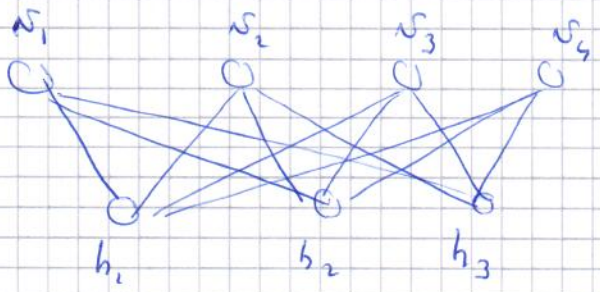
2). Restricted Boltzman Machine (RBM).

In the learning theory community the Ising Model is also (abusively) called a "Boltzman machine".

There also exist a variant that is popular with many applications (e.g as a generative model for pdf's) called the Restricted Boltzman Machine (RBM).

$$G = (\underbrace{\text{Visible} \cup \text{Hidden}}_{\text{vertices}}, \underbrace{E}_{\substack{\text{edges between} \\ \text{visible \& hidden only}}})$$

a "complete bipartite" graph:



visible units

hidden units

$$p(\underline{v}, \underline{h}) = \frac{1}{Z} \exp\left\{ - \sum_{\substack{i \in Vis \\ j \in Hid}} W_{ij} v_i h_j - \sum_{i \in Vis} b_i v_i - \sum_{j \in Hid} c_j h_j \right\}$$

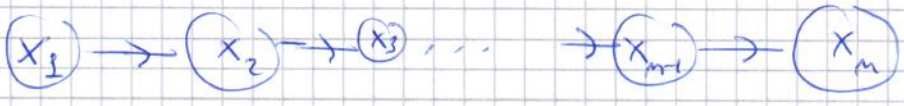
Here $\underbrace{\max}$ cliques are pairs (i,j).

$p_{target}(\underline{v}) = \sum_{\underline{h} \in Hidden} p(\underline{v}, \underline{h})$, learn W_{ij}, b_i, c_j from sample $(\underline{v}^{(1)}, \dots, \underline{v}^{(N)})$ of visible variables $\sim p_{target}(\underline{v})$

IV Conversion of BN into MRF.

This conversion is useful often. Here we illustrate with two examples and give the general algorithm later on.

a) Simple example of a chain graph (e.g. a Markov process).



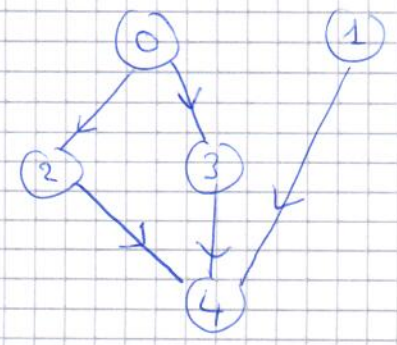
$$p(\underline{x}) = \underbrace{p(x_1)\, p(x_2|x_1)}_{\psi_{12}(x_1,x_2)} \underbrace{p(x_3|x_2)}_{\psi_{23}(x_2,x_3)} \cdots \underbrace{p(x_n|x_{n-1})}_{\psi(x_n,x_{n-1})}$$

$$= \frac{1}{Z}\, \psi_{12}(x_1,x_2) \cdots \psi(x_n, x_{n-1}) .$$

with $Z = 1$ here (recall BN are automatically normalized!).

Cliques are $(12), (23), \ldots, (n-1, n)$. The MRF is the same without arrows (it is undirected)

b) Generic example.



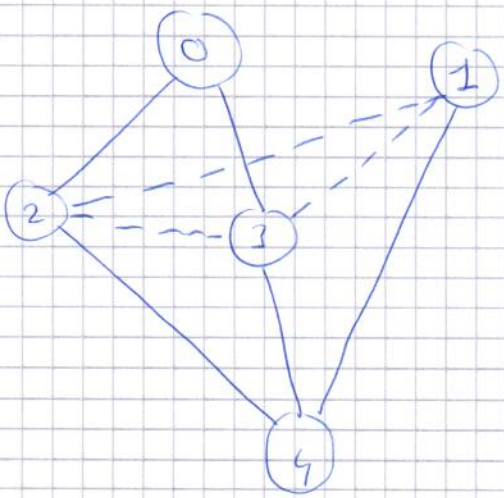
$$p(\underline{x}) = p(x_0)\, p(x_1)\, p(x_2|x_0)\, p(x_3|x_0)$$
$$\cdot\, p(x_4|\, x_2\, x_3\, x_1)$$

$$p(\underline{x}) = \underbrace{p(x_0)\, p(x_2 \mid x_0)\, p(x_3 \mid x_0)}\quad \underbrace{p(x_4 \mid x_2 x_3 x_1)\, p(x_1)}$$

$$= \frac{1}{Z} \quad \psi_{023}(x_0 x_2 x_3) \quad \psi_{1234}(x_1 x_2 x_3 x_4)$$

with $Z = 1$.

MRF graph is:



- added edges between parents of childs
- This process of "adding edges between parents of a child" is called the "moralization step".
- The MRF graph obtained from the BN DAG is called the "Moralization graph".

c) Conversion Algorithm BN → MRF.

- Take a DAG.
- For each Node connect all its parents and remove arrows. We get the Moralization graph.
- Take each Max Clique in Moralization graph and define $\psi_c(x_c)$ by the product of cond prob distr appearing in DAG of each node in the clique. (see example b) above)