① Lecture 5 - Markov Random Fields (MRFs)

So far we've discussed 1) pdfs $p(x|y,z)$ (but not multivariate, e.g. $p(\underline{X}|Y,Z)$)

2) directed graphic models (Bayes nets)

Are there other ways to represent complex distributions (which is to say <u>knowledge</u>)

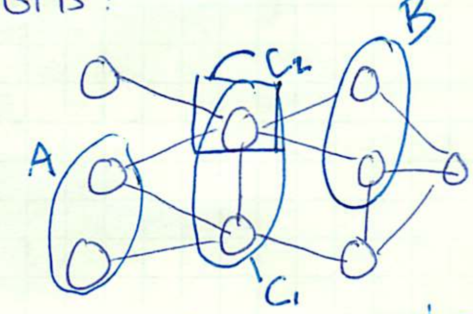
Bayes Nets (or DAGs) are one way to factor (and therefore simplify a complex joint prob distr.

$$P(x_{1:N}) = \prod_{i=1}^{N} p(x_i | pa(x_i))$$

$$p(x_i | pa(x_i)) = p(x_i) \quad \text{if } pa(x_i) = \emptyset .$$

Are there other ways? Yes.

Undirected GMs:



nodes still represent variables, but now there is no causality implied, i.e. no arrows.

what are the independence properties of this graph like?

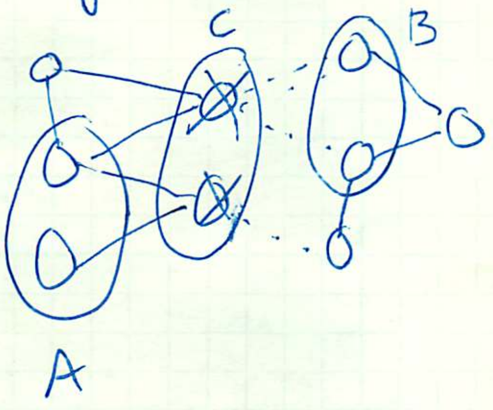
$A \perp\!\!\!\perp B | C$?

Much simpler than in DAGs: If all paths from A to B pass through C, the C "blocks" A & B and $A \perp\!\!\!\perp B | C_1$.

But if there is a path that isn't blocked the independ. cond. no longer holds $A \not\perp\!\!\!\perp B | C_2$

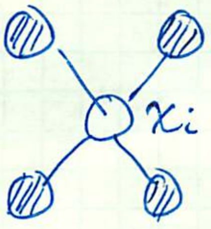
There is no "explaining away" phenom.

Another way to think about it.



If we remove nodes in C from graph, and there is no path from A to B then $A \perp\!\!\!\perp B \mid C$.

- Markov blanket is simpler too.

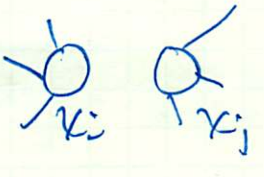


~~x_i only depends on its~~ ~~neighboring~~

- x_i is cond. indep of all other nodes given only its neighbors.

- What factorization model do we use?

What does a ~~[illegible]~~ link mean? (or absense of one?



If ~~[illegible]~~ the is no connection between nodes x_i & x_j then ~~[illegible]~~ $x_i \perp\!\!\!\perp x_j \mid$ rest of graph. because all other paths are blocked.

$p(x_i, x_j \mid x_{k \neq i|j})$

or $x_{k \setminus \{i,j\}}$

$= p(x_i \mid x_{k \neq i,j})\, p(x_j \mid x_{k \neq i,j})$

491

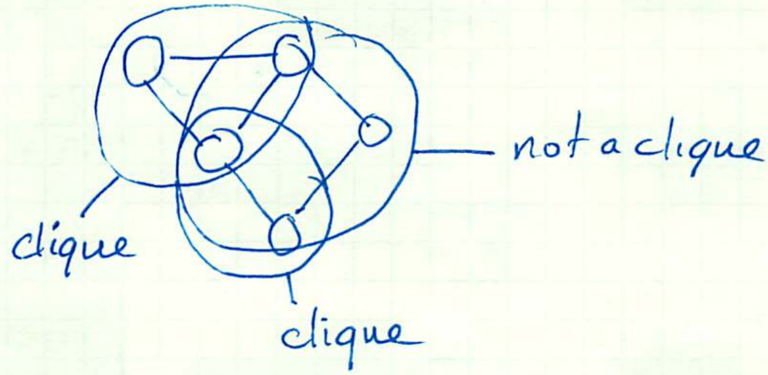
3 L5 MRFs

Before we define joint prob. we need to introduce concept of cliques

Cliques of a graph:

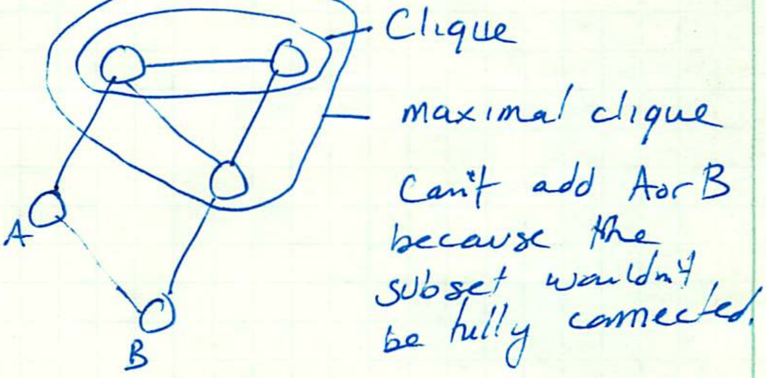
a subset of nodes s.t. ∃ a link between all pairs of nodes in the subset.

The set of nodes ~~that~~ in a clique is fully connected



Maximal cliques:

A clique where it is not possible to add any more nodes without breaking the clique, i.e. by including any more nodes the subset will no longer be fully connected and cease to be a clique:



Can't add A or B because the subset wouldn't be fully connected.

Define joint distribution to be functions of maximal cliques — ~~any~~ using any subset of a maximal clique would be redundant.

Let C denote a clique

x_c = set of vars in C.

$$p(\overset{=\underline{x}}{x_{1:N}}) = \frac{1}{Z} \prod_C \psi_c(\underline{x}_c) \quad , \quad \psi_c(\underline{x}_c) \geq 0$$

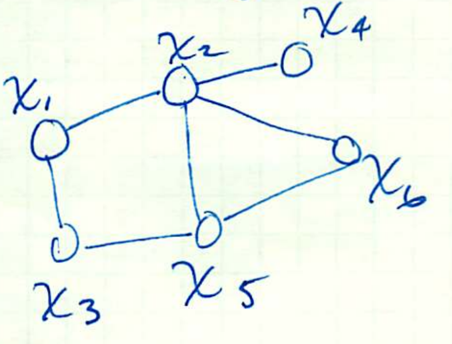
This factorizes the joint prob. distri into small joint pdfs defined on cliques.

Z = normalization const to make $p(x_{1:N})$ a valid pdf. Also called <u>partition function</u>.

$$Z = \sum_{\underline{x}} \prod_c \psi_c(\underline{x}_c)$$

(assume discrete vars, but could use continuous vars and integrals

$$Z = \int \prod_c \psi_c(\underline{x}_c)\, d\underline{x}_c$$



$$p(\underline{x}) = \frac{1}{Z} \psi(x_1, x_2)\psi(x_1, x_3)\psi(x_2, x_4)$$
$$\cdot \psi(x_3, x_5) \cdot \psi(x_2, x_5, x_6)$$

This is also valid, don't have to restrict pdf to maximal cliques:

$$p(\underline{x}) = \frac{1}{Z} \psi(x_1, x_2)\psi_{13} \cdot \psi_{25}\psi_{24}\psi_{25}\psi_{26}\psi_{56} \quad \text{why?}$$

Equivalent ~~to~~ assuming $\psi_{256} = \psi_{25}\psi_{26}\psi_{56}$, i.e. it factorizes.

What are "potential functions"? Are they pdfs?

No (or not necessarily)

Only need $\Psi(\underline{x}) \geq 0$ so that $p(\underline{x}) \geq 0$

$\Psi(x)$ does not have a specific interpretation like conditional or marginal, unlike DAGs.

$\boxed{\text{Problem:}}$ does not produce a normalized joint pdf.

<u>Need</u> Z, which is often hard to compute.

↳ can be a major limitation of UDGs

How bad is it? How big is the sum?

$$Z = \sum_{\underline{x}} \prod_{c} \Psi_c(\underline{x}_c)$$

↳ this sums over all values of $\underline{x}$.

- Typically M nodes, each with K states
  $\Rightarrow$ K^M different states!

- Need Z for parameter learning:

$$\frac{\partial \Psi_c(x_c|\theta_c)\frac{1}{Z}}{\partial \theta_c}$$

- But not for local cond. probs:

$$P(\underline{x}_a|\underline{x}_b) = \frac{\cancel{\tfrac{1}{Z}} P(\underline{x}_a, \underline{x}_b)}{\cancel{\tfrac{1}{Z}} P(\underline{x}_b)} \qquad Z\text{'s cancel}$$

- There are some techniques which we will come across later.

Hammersley-Clifford theorem : Relating factorization to cond. indep.

#1) Undirected GM G is an MRF if two nodes are conditionally indep whenever they are separated by evidence nodes.

$$P(X_i | X_{G\setminus i}) = P(X_i | X_{N_i})$$

$G\setminus i$ = all nodes in graph G except i

N_i = neighboring nodes of X_i

#2) a pdf $P(X)$ on an undirected GM is is a Gibbs distribution if it can be factored into positive functions defined on cliques that cover all nodes and edges of G.

$$P(X) = \frac{1}{Z} \prod_{C \in C_G} \psi_C(X_C)$$

C_G = all (maximal) cliques

H-C theorem says ~~Def(1) ⟺ Def 2~~

Def #1 ⟺ Def #2

$\psi_c(\underline{x}_c) > 0$ so it's convenient to write

$$\psi_c(\underline{x}_c) = \exp[-E(\underline{x}_c)]$$

$$\boxed{E(\underline{x}_c) = \text{"energy function"}}$$

as in "energy-based" models

Exponential representation =

$$\boxed{\exp[-E(\underline{x}_c)] = \text{"Boltzman distribution"}}$$

Total energy is the sum of energies in each clique.

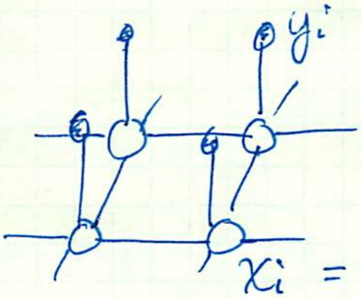
- potentials (and energy functions) do not have a specific probabilistic interpretation.

  So, what are they? How do we choose them?

- Potential fns express "good" configurations of local vars.

Example: image denoising with binary pixels

$y_i \in \{-1, +1\}$



y_i = noisy pixels (observed)
10% pixels flipped.

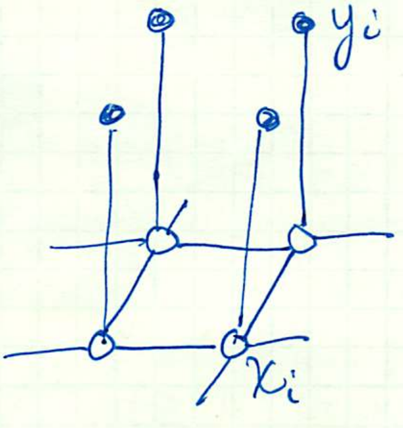
x_i = noise-free true pixel values (hidden)

How do we express knowledge in the network?

Eg. we know pixels are correlated.

→ links specify degree of correlation.

How do we set the energy functions?



x_i & y_i should be correlated

can use $-\eta x_i y_i$

same ⇒ low energy (good)

diff ⇒ high energy

η = const > 0

The rest of the graph:

$-\beta x_i x_j$ same ⇒ lower energy

$\beta = 0$ ⇒ no links diff ⇒ higher energy β = const > 0

Can also model tendency for pixels to be on or off:

$h x_i$ h = const > 0 "bias" h = 0 ⇒ equal prob +1, -1

Still valid because only cond. is for energy fn to be > 0.

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

Just add energies of all cliques ⇒

$$p(\underline{x}, y) = \frac{1}{Z} \exp\left[-E(\underline{x}, y)\right]$$

We are given observed pixels, so we have $p(\underline{x} | y)$ defined implicitly

How do we obain x_i?

Start at some soln, change x_i to maximize energy.

iterated conditional modes (ICM)

coord-wise gradient ascent.

1) $x_i = y_i \; \forall i$

2) for each x_i, (or at random)

calc $E \mid x_i = +1$

$E \mid x_i = -1$

everything else is held fixed.

3) change x_i to state with lower energy.

can do efficiently because only one term in $E(\underline{x})$ changes.

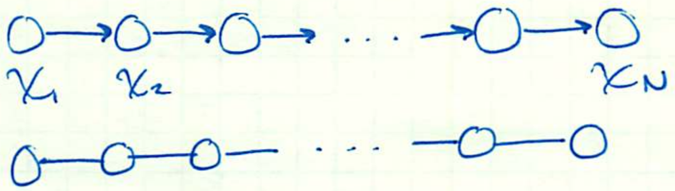
4) repeat until stable (or run "long enough")

will converge to local maximum of $E(\underline{x})$ (not global)

[see slides for example]

How do these related to directed graphs?
Can we convert one to another?
Are there adv. / disadv. to each?

Consider:



For DAG $p(x) =$ ~~[illegible]~~

$$p(x_1)p(x_2|x_1)p(x_3|x_2)\cdots p(x_N|x_{N-1})$$

For UDG:

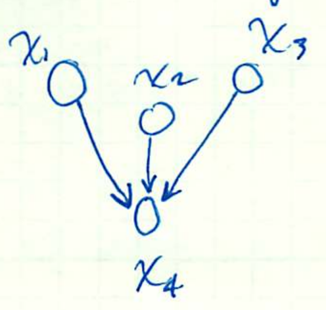
$$p(x) = \frac{1}{Z}\psi_{1,2}(x_1,x_2)\psi_{2,3}(x_2,x_3)\ldots\psi_{N-1,N}()$$

$\Rightarrow$ can set $\psi_{1,2}(x_1,x_2) = p(x_1)p(x_2|x_1)$

etc.

$\Rightarrow Z=1$ since it's already normalized.

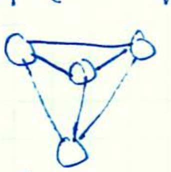
Can we do this in general?



$$p(\underline{x}) = p(x_1)p(x_2)p(x_3)p(x_4|x_{1:3})$$

Can we use the same trick?

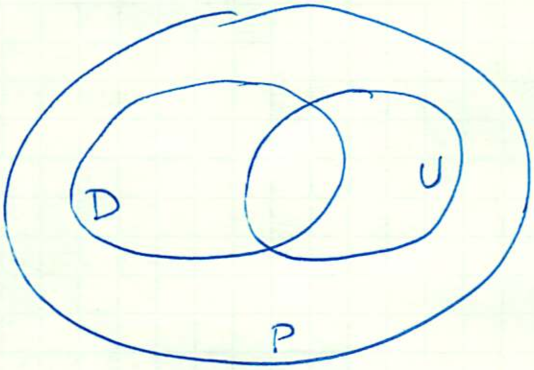
$$p_{UG}(x) = p(x_1)p(x_2)p(x_3)\underbrace{p(x_4|x_1,x_2,x_3)}$$

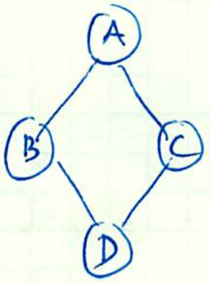


but this is fully connected $\Rightarrow$ no cond indep properties, no adv.

This involves all 4.
$\Rightarrow$ ~~[illegible]~~ all belong to a single clique

UGs vs DGs

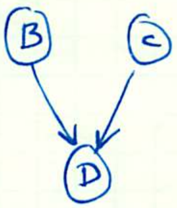




No DG can represent only

$A \perp D \mid B, C$

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



No UG can rep. only

$B \perp C \mid \emptyset$