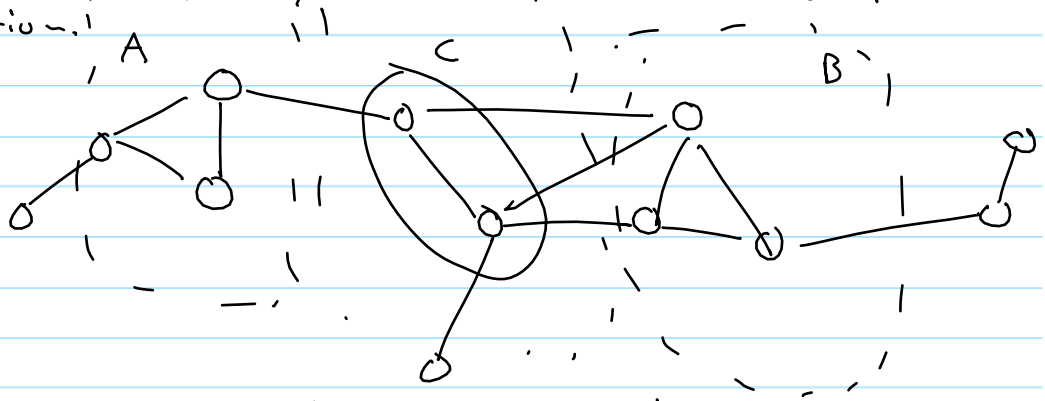
Markov Random Fields

- Described by undirected graphical models
- Major class of graphical models
- Also specify factorization & conditional independences
  * though differently

MRF ~~factorization~~ - corresponding to graph separation.



To test if $A \perp B \mid C$ we look at all possible paths from all nodes in A to all nodes in B. If all pass through one or more nodes in C then such paths are blocked.

- No explaining away
- testing conditional independence simpler

- Factorization properties
  - if x_i and x_j are not connected by a link
    - $p(x_i, x_j \mid \vec{x}_{\setminus\{i,j\}}) = p(x_i \mid \vec{x}_{\setminus\{i,j\}}) p(x_j \mid \vec{x}_{\setminus\{i,j\}})$

    because there is no direct path and all other nodes observed

How do we parameterize an MRF?

- Clique potentials, all other parameterizations would be redundant

Let C be a clique and the var's in that clique are $\vec{x}_c$, then the joint dist. is written as a product of pot. fnacts $\psi_c(\vec{x}_c)$ over the max cliques of the graph.

$$p(\vec{x}) = \frac{1}{Z} \prod_c \psi_c(\vec{x}_c)$$

where Z, the "partition function" is a normalizing co-stant

$$Z = \sum_x \prod_c \psi_c(\vec{x}_c)$$

we will consider only potential function $\psi_c(x_c) \geq 0$

- Note $\psi_c(x_c)$ need not be a prob. function!

- Z is a problem, in model with M discrete K state nodes $\sum_x$ involves K^M states

- partition function is usually needed for learning because it will be a function of whatever parameters the potential functions have.

- to compute local conditionals the partition func is not needed because the cond. is a ratio of terms that cancel.

- local marginals can be locally normalized

Conditional independence

Requires $\psi_c(x_i) > 0$
↑ not ge

Beyond scope for the class. the Hammersley-Clifford theorem states that the set of distributions that can be factorized into the product of maximal cliques is the same as the set of distributions whose conditional independence structure can be read off of the graph separability.

Often potentials are expressed as exponentials ie.

$$\psi_c(\vec{x}_c) = \exp(-E(x_c)) \quad \text{where}$$

$E(\vec{x}_c)$ is called an energy function.

Example application: Image denoising

Binary images $y_i \in \{-1, +1\}$ pixels observed $i = 1..D$

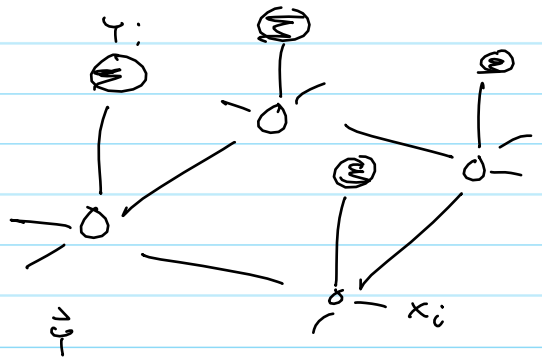
$x_i \in \{-1, +1\}$ noisy observed image

noise = flipping a bit w.p. 10%

assume strong correlation btwn x_i & y_i and between neighboring pixels x_i & x_j

$$E(\vec{x}, \vec{y}) = h \underbrace{\sum_i x_i}_{\text{bias}} - \beta \underbrace{\sum_{\{i,j\}} x_i x_j}_{\text{agreement btw neighbors}} - \eta \underbrace{\sum x_i y_i}_{\text{agreement w/ observation}}$$

- explain signs



if we fix $\vec{y}$

this defines a conditional dist. $p(\vec{x}|\vec{y})$ over noise-free images. A goal is to find an $\vec{x}$ that has high probability.

One approach, Iterated Conditional Modes
- basically coordinate-wise gradient ascent
- initialize all $x_i := y_i$, evaluate
- conditionals for single x_i and set x_i to one with lowest energy
- repeat

Relationship to directed graphs

Directed graph → undirected

example

$$p(\vec{x}) = p(x_1)\,p(x_2|x_1) \cdots p(x_N|x_{N-1})$$

to convert this to max cliques simply write

$$p(\vec{x}) = \frac{1}{Z}\,\psi'_{1,2}(x_1,x_2)\,\psi_{2,3}(x_2,x_3)\cdots\psi_{x_{n-1}x_n}(x_{n-1},x_n)$$

where

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

$$\psi_{i,j}(x_i,x_j) = p(x_j|x_i) \quad \forall \{i,j\} \neq \{1,2\}$$

and

$$Z = 1$$

Moralization

More generally this conversion requires "marrying the parents". In this chain example the moral graph is complete.

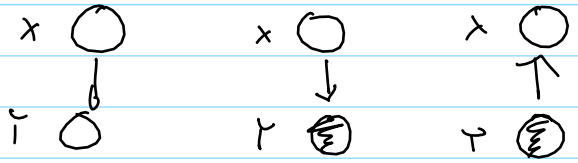
Recipe: directed G.M → undirected G.M.

1) Add links between all pairs of parents for all nodes in graph
2) Drop arrows
3) Initial clique potentials to 1
   a) Multiply in all conditional dist's associated with each clique.
4) $Z = 1$

* Inference in Graphical Models * ← KEY

Idea: exploit graphical structure in algorithms for inference.

First graphical Bayes Theorem



Joint $p(x, y) = p(x) \cdot p(y|x)$

If we obs. y then $p(x)$ can be seen as a prior on x, and inferring the post. dist. of x can be the goal. To do this note

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

and $p(x|y) = \dfrac{p(y|x)\, p(x)}{p(y)}$, reversing the arrow