Lecture 6: Markov random fields

- Structure of a Markov random field
- Potentials
- Relationship to directed models

Undirected graphical models

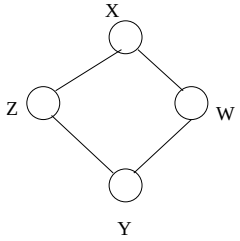
- So far we have used directed graphs as the underlying structure of a Bayes net
- Why not use *undirected* graphs as well?
  E.g., variables might not be in a "causality" relation, but they can still be correlated, like the pixels in a neighborhood in an image
- An undirected graph over a set of random variables $\{X_1, \ldots X_n\}$ is called a **undirected graphical model** or **Markov random field** or **Markov network**

Conditional independence

- We need to be able to specify, for a given graph, if $X \perp\!\!\!\perp Z | Y$, for any disjoint subsets of nodes X, Y, Z.
- In directed graphs, we did this using the Bayes Ball algorithm
- In undirected graphs, independence can be established by a simple notion of separation: if every path from a node in X to a node in Z goes through a node in Y, we conclude that $X \perp\!\!\!\perp Z | Y$
- Hence, independence can be established by removing the nodes in the conditioning set then doing reachability on the remaining graph.
- What is the Markov blanket of a node in an undirected model?

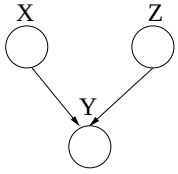
How expressive are undirected models?

- Example 1: an undirected graph



  Can we find a directed graph that satisfies the same independence relations?

- Example 2: our friend the v-structure



  Can we find an undirected graph that satisfies the same independence relations?

Local parameterization

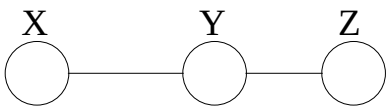
- In directed models, we had local probability models (CPDs) attached to every node, giving the conditional probability of the corresponding random variable given its parents
- Can we do something similar in undirected models?
- More specifically, we want the joint probability distribution to factorize over the graph
- This means that the joint can be written as a product of "local" factors, which depend on subsets of the variables.
- Unfortunately, conditional probabilities are not adequate for this case...

What about local marginal parameterizations?

- Suppose we express the joint as:

$$p(X_1, \ldots X_n) = \prod_i p(X_i, \text{Neighbors}(X_i))$$

- It is local and has a nice interpretation
- So let's consider using it for an example:



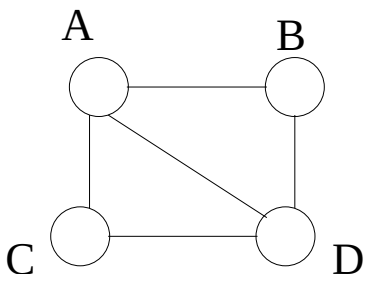
Local parameterizations: Try 2

- Consider a pair of nodes X and Y that are not directly connected through an arc
- According to the conditional independence interpretation, X and Y are independent given all the other nodes in the graph

$$X \perp\!\!\!\perp Y | (\{X_1, \ldots X_n\} - X - Y$$

- Hence, there must be a factorization in which they do not appear in the same factor
- This suggests that we should define factors on **cliques**
  Recall that a clique is a fully connected subset of nodes (i.e., there is an arc between every pair of nodes)

Example: what are the cliques?



Clique potentials

- We will represent the joint distribution as a **product of clique potentials**:

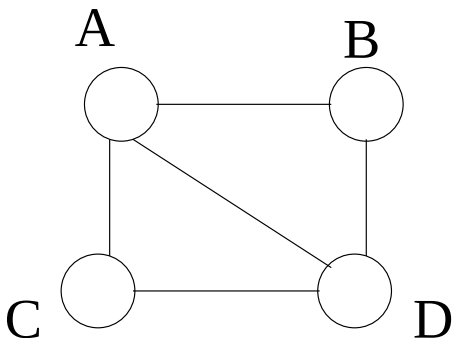
$$p(X_1, \ldots X_n) = \frac{1}{Z} \prod_{\text{cliques } C} \psi_C(\mathbf{x}_C)$$

where $\mathbf{x}_C$ are values assignments for the variables that participate in the clique and Z is a normalization constant, to make probabilities sum to 1:

$$Z = \sum_{\mathbf{x}} \prod_{\text{cliques } C} \psi_C(\mathbf{x}_C)$$

- Without loss of generality, we can consider only *maximal cliques* These are the cliques that cannot be extended with other nodes without losing the fully connected property

Example

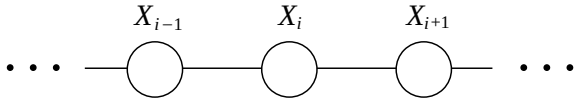


Normalizing constant

- The normalizing constant Z can be ugly to compute, since we have to sum over all possible assignments of values to variables
- Depending on the shape of the graph, the summation could be done efficiently
- However, if we are interested in conditional probabilities, we do not even need to compute it! (why?)

Interpretation of clique potentials

- Potentials are *NOT* probabilities (conditional or marginal)
- But they do have a natural interpretation as “agreement” or “energy”
- Example: spin glass model



(a)

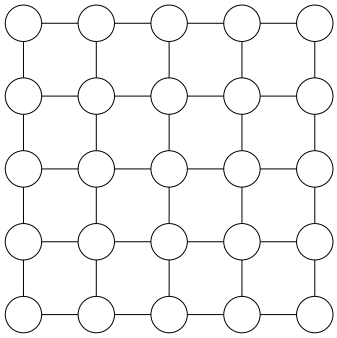
x_{i-1} \ x_i	–1	1
–1	1.5	0.2
1	0.2	1.5

x_i \ x_{i+1}	–1	1
–1	1.5	0.2
1	0.2	1.5

(b)

A real example: Texture synthesis

- You are given a small patch of texture and want to produce a "similar" larger patch
- We can define a Markov random field over pixels, e.g:



- The "potentials" favor certain configurations of pixels over others
- We get the texture by doing inference (and sometimes learning) for this model

Boltzmann distribution

- The fact that potentials must be non-negative is annoying
- We can escape from that by using the exponential function, which is non-negative:

$$\psi_C(\mathbf{x}_C) = e^{-H_C(\mathbf{x}_C)}$$

- Now we have to define $H_C(\mathbf{x}_C)$, which can be anything!
- Moreover, the joint also has a nice form:

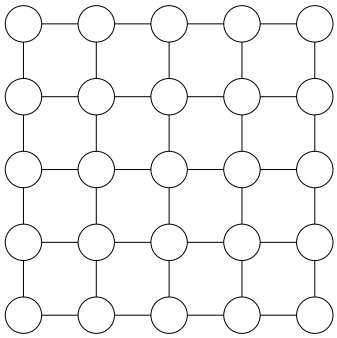
$$p(X) = \frac{1}{Z}\prod_C e^{-H_C(\mathbf{x}_C)} = \frac{1}{Z}e^{-\sum_C H_C(\mathbf{x}_C)} = \frac{1}{Z}e^{-H(X)}$$

where $H(X) = \sum_C H_C(\mathbf{x}_C)$ is the "free energy"

- Hence, p is represented using a *Boltzmann distribution*

Ising Model

- Nodes are arranged in a regular fashion and connected only to geometric neighbors.
- E.g., Spin glass in 2D:



- Energy has the form"

$$H(X) = \sum_{i,j} \beta_{ij} x_i x_j + \sum_i \alpha_i x_i$$

Important result

- Consider the family of probability distributions that respect all the conditional independencies implied by an undirected graph G
- Consider the family of probability distributions defined by ranging over all possible maximal clique potential functions.
- The Hammersley-Clifford theorem shows that these two families are identical.
- This is a similar result to the "soundness and completeness" of d-separation which we discussed for directed models.