1. How do we build, represent and work with joint probability distributions over a large number of variables?

   (a) Our main tool is to exploit conditional independences.
   
   (b) Graphical models, in general, are a way to represent the conditional independences of a joint probability distribution in a compact way, by representing them in a graph.
   
   i. The nodes represent the variables.
   ii. The edges represent something about the dependency among the variables.
   
   (c) In class, we will talk about Directed graphical models, where the edges in the graph are directed, which try to capture the causal dependence of one variable on another.
   
   (d) Today, we will talk about Undirected graphical models, where the edges are undirected.

2. Undirected Graphical Models

   (a) Alternative names: Markov Random Field (MRF) or Markov network.
   
   (b) How do UGMs represent conditional independencies:
   
   i. **Global Markov Property**: for sets of nodes $A, B, C,$

      $$x_A \perp x_B \mid x_C$$

      if $C$ separates $A$ from $B$ in the graph. $x_A$ refers to the variables in set $A$.
   
   ii. Note vice versa. The graph tells us which CIs *must* exist, not the other way around.

---

![Example UGM](image-url)
iii. One way to think of a graph is that it specifies a set of join probability distributions, namely, those that satisfy that conditional inddependencies implied by the graph separation.

(c) **Markov Blanket:** The markov blanket of a node \( t \) is defined as the set of nodes \( mb(t) \) that renders \( t \) conditionally independent of the rest of the nodes in the graph given \( mb(t) \). By the graph separation property, the markov blanket of a node of \( t \) is the set of \( t \)'s immediate neighbors.

(d) **pairwise Markov property:** if there is no edge between two nodes, then they are conditionally independent given the rest of the graph

\[
s \propto t \mid V \setminus s, t
\]

(e) pairwise Markov property implies global Markov property and vice versa. We won’t go over the proof here.

(f) **Expressiveness:** UGMS cannot represent the conditional independence of every probability distribution.

   i. Remember that a UGM specifies which CIs definitely exist. There can be additional ones it doesn’t capture.

   ii. For example: consider a distribution over variables \( a, b, c \) where we sample \( a \) and \( b \) from independent prior distributions and \( c \) depends on both \( a \) and \( c \). So \( p(a, b, c) = p(a) p(b) p(c | a, b) \).

   A. \( a \) is unconditionally independent of \( b \), so there should not be a path between \( a \) and \( b \).

   B. but both \( a \) and \( b \) must be connected \( c \), so there will always be a path between \( a \) and \( b \).

(g) **Parameterization of UGMS:**

   i. **Potential functions:** we associate a potential function \( \psi_c(y_c) \) with every clique of in graph, where \( psi_i \) is any function that assigns a non-negative value to any assignment of values of the variables in clique \( c \).

   ii. We then write \( p(y) \propto \prod_{c \in C} \psi_c(y_c) \).

   iii. By the **Hammersly-Clifford** theorem: a positive distribution \( p(y) > 0 \) satisfies the CI properties of an undirected graph \( G \) iff \( p \) can be represented as a products of potentials, one per clique of \( G \).

   iv. **Gibbs distribution:** By the H-C theorem, we are free to assign the following distribution to a graph \( G \):

\[
p(y | \theta) = \frac{1}{Z(\gamma)} \exp(- \sum_c E(y_c | \gamma))
\]

. This is called the Gibbs distribution. Here, \( E \) refers to an energy function which corresponds to the compatibility for the variable assignments.

A. **Partition Function:** \( Z \), the normalizing constant which is a function of \( \theta \).
v. **Pairwise MRFs:** It is often simplest to assume that the probability distribution can be factorized into pairwise potentials:

A. \[ p(y|\theta) \propto \prod_{e,ij \in E} \psi_{ij}(y_i, y_j) \]

B. This restricts the probability distributions in our parameterization more.

(h) How do we represent potentials functions?

i. **Maximum-Entropy or Log-Linear:** in this case we say the value of a potential on an input is a linear combination of some features of the input:

\[ \log p(y|\theta) = \sum_c \phi_c(y_c)^T \theta_c - Z(\theta) \]

(i) Example MRFs:

i. **Ising Model:** Binary variables \( y_i \in \{-1, 1\} \) arranged in a lattice (say, 2-dimensional), where the potentials are pairwise and symmetric and \( \psi(1, 1) = \psi(-1, -1) = e^J \) and \( \psi(1, -1) = \psi(-1, 1) = e^{-J} \).

(j) if \( J > 0 \) then we have two modes in which all the variables are the same.

(k) if \( J < 0 \) then all the variables want to be different and we have a much more complex system.

i. **Gaussian MRF:** Each node and each edge is associated with a gaussian distribution.

(l) It turns out that if we write the join of this distribution in *information form*:

\[ p(y|\theta) \propto \exp[\eta^T y - \frac{1}{2} y^T \Lambda y] \]