

Graphical Models

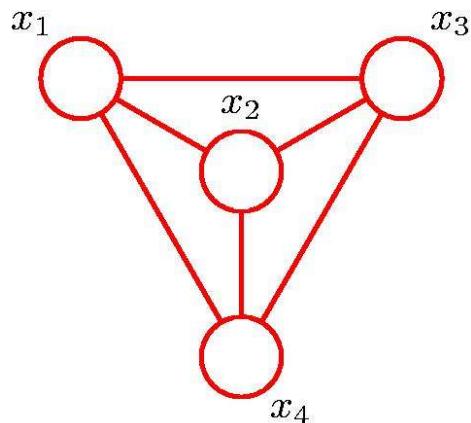
Vahid Tarokh
ECE 685D, Fall 2025

Graphical Models

- Probabilistic graphical models provide a powerful framework for representing **dependency structure between random variables**.
- Graphical models offer several useful properties:
 - They provide a simple way to visualize the structure of a probabilistic model and can be used to motivate new models.
 - They provide **various insights into the properties of the model**, including conditional independence.
 - Complex computations (e.g. inference and learning in sophisticated models) can be expressed in terms of graphical manipulations.

Graphical Models

- A graph contains a set of nodes (vertices) connected by links (edges or arcs)



- In a probabilistic graphical model, each node represents a random variable, and links represent probabilistic dependencies between random variables.
- The graph specifies the way in which the joint distribution over all random variables decomposes into a product of factors, where each factor depends on a subset of the variables.

- Types of graphical models:

- Bayesian networks, also known as Directed Graphical Models (the links have a particular directionality indicated by the arrows)
- Markov Random Fields, also known as Undirected Graphical Models (the links do not carry arrows and have no directional significance).
- Hybrid graphical models that combine directed and undirected graphical models, such as Deep Belief Networks.

Bayesian Networks

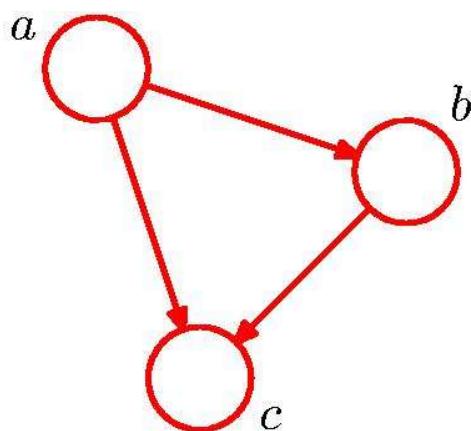
- Directed Graphs are useful for expressing **causal relationships** between random variables.
- We consider an arbitrary joint distribution $p(a, b, c)$ over three random variables a, b and c.
 - Note that at this point, we do not need to specify anything else about these variables
 - By application of the product rule of probability (twice):
$$p(a, b, c) = p(c|a, b)p(a, b) = p(c|a, b)p(b|a)p(a)$$
- This decomposition holds for any choice of the joint distribution.

Bayesian Networks

- By application of the product rule of probability (twice), we get

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

- Represent the joint distribution in terms of a simple graphical model:



- Introduce a node for each of the random variables.
- Associate each node with the corresponding conditional distribution in above equation.
- For each conditional distribution we add directed links to the graph from the nodes corresponding to the variables on which the distribution is conditioned.

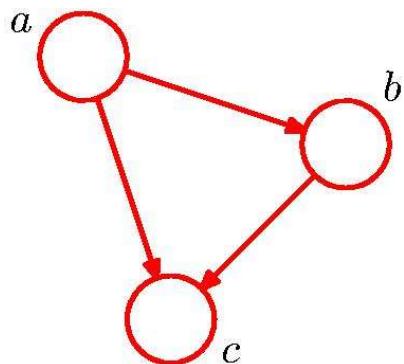
- Hence for the factor $p(c|a, b)$, there will be links from nodes a and b to node c.
- For the factor $p(a)$, there will be no incoming links.

Bayesian Networks

- By application of the product rule of probability (twice), we get

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

- If there is a link going from node a to node b, then we say that:



- node a is a **parent** of node b.
- node b is a **child** of node a.

- For the decomposition, we choose **a specific ordering** of the random variables: a,b,c.
 - If we chose a **different ordering**, we would get a **different graphical representation** (we will come back to that point later).

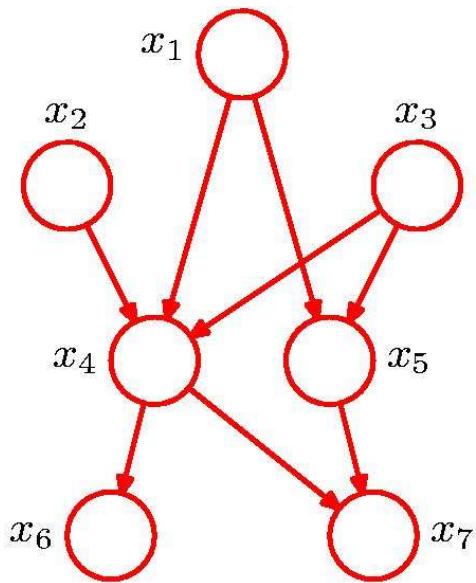
- The joint distribution over K variables factorizes:

$$p(x_1, \dots, x_K) = p(x_K|x_1, \dots, x_{K-1}) \dots p(x_2|x_1)p(x_1)$$

- If each node has incoming links from all lower numbered nodes, then the graph is **fully connected**; there is a link between all pairs of nodes.

Bayesian Networks

- Absence of links conveys certain information about the properties of the class of distributions that the graph conveys.



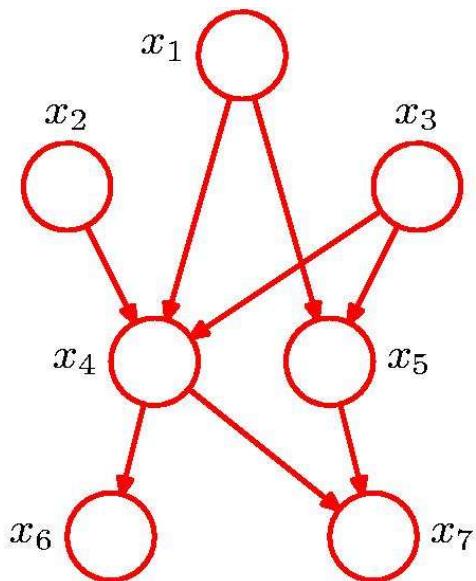
- Note that this graph is not fully connected (e.g. there is no link from x_1 to x_2).
- The joint distribution over x_1, \dots, x_7 can be written as **a product of a set of conditional distributions**.

$$\begin{aligned} p(x_1, \dots, x_7) &= p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3) \\ &\quad p(x_5|x_1, x_3)p(x_6|x_4)p(x_7|x_4, x_5) \end{aligned}$$

- Note that according to the graph, x_5 will be conditioned only on x_1 and x_3 .

Factorization Property

- The joint distribution defined by the graph is given by **the product of a conditional distribution** for each node conditioned on its parents:



$$p(\mathbf{x}) = \prod_{k=1}^K p(x_k | \text{pa}_k)$$

- Where pa_k denotes a set of parents for the node x_k .
- This equation expresses a **key factorization property** of the joint distribution for a **directed** graphical model.
- Important restriction: There must be **no directed cycles!**
- Such graphs are also called **directed acyclic graphs (DAGs)**.

Bayesian Curve Fitting

- As an example, consider Bayesian polynomial regression model:

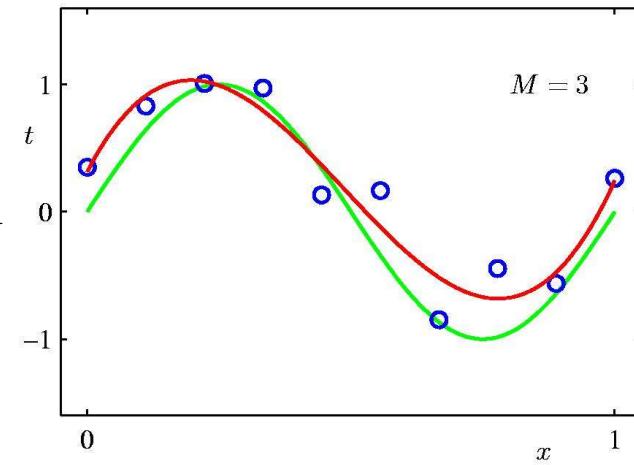
$$y(x, \mathbf{w}) = \sum_{j=0}^M w_j x^j$$

- We are given inputs $\mathbf{X} = \{x_1, x_2, \dots, x_N\}$ and target values $\mathbf{t} = [t_1, t_2, \dots, t_N]^T$.

- Given the prior over parameters, the joint distribution is given by:

$$p(\mathbf{t}, \mathbf{w} | \mathbf{X}) = p(\mathbf{w}) \prod_{n=1}^N p(t_n | y(\mathbf{w}, x_n)).$$

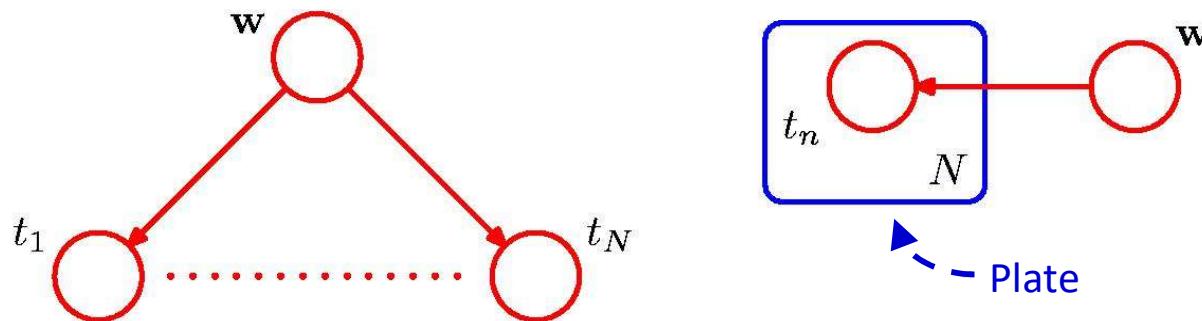
↑ ⌈ ⌉
Prior term Likelihood term



Graphical Representation

$$p(\mathbf{t}, \mathbf{w} | \mathbf{X}) = p(\mathbf{w}) \prod_{n=1}^N p(t_n | y(\mathbf{w}, x_n)).$$

- This distribution can be represented as a graphical model.
- Same representation using plate notation.

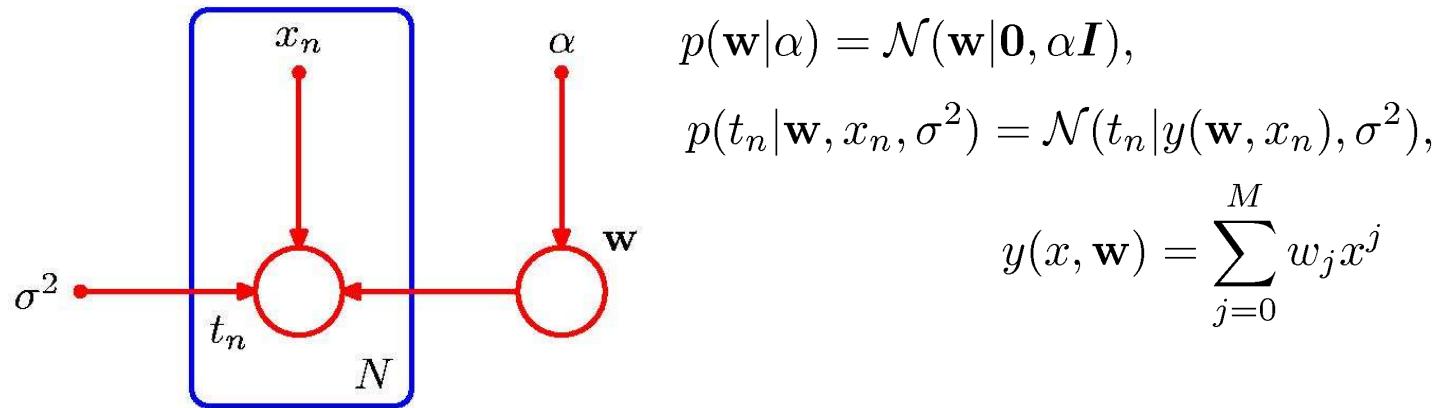


- **Compact representation:** we introduce a plate that represents N nodes of which only a single example t_n is shown explicitly.
- Note that w and $\mathbf{t} = [t_1, t_2, \dots, t_N]^T$ represent random variables.

Graphical Representation

- It will often be useful to make the parameters of the model as well as random variables be explicit.

$$p(\mathbf{t}, \mathbf{w} | \mathbf{x}, \alpha, \sigma^2) = p(\mathbf{w} | \alpha) \prod_{n=1}^N p(t_n | \mathbf{w}, x_n, \sigma^2).$$



- Random variables will be denoted by **open circles** and deterministic parameters will be denoted by **smaller solid circles**.

Graphical Representation

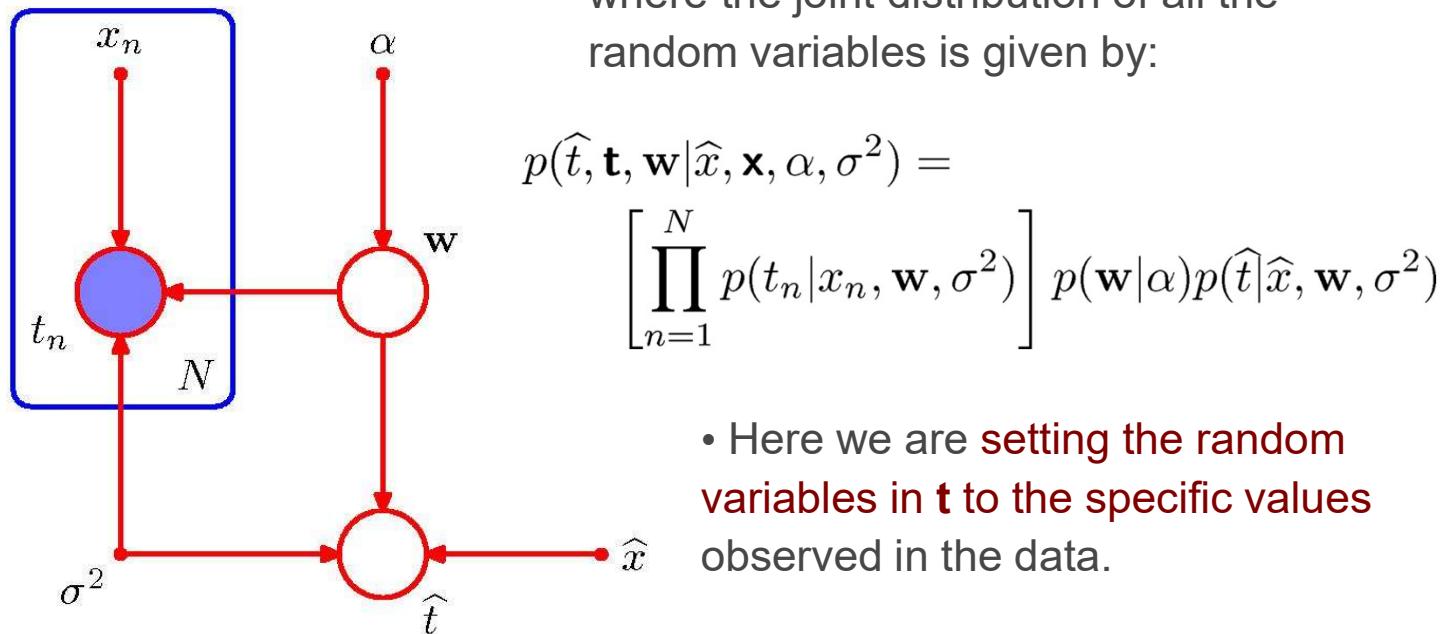
- When we apply a graphical model to a problem in machine learning, we will set some of the variables to specific observed values (e.g. condition on the data).
 - For example, having observed the values of the targets $\{t_n\}$ on the training data, we wish to infer **the posterior distribution over parameters w .**
 - In this example, we conditioned on observed data $\mathbf{t} = [t_1, t_2, \dots, t_N]^T$ by shadowing the corresponding nodes.
-
- $$p(\mathbf{w}|\mathbf{t}) \propto p(\mathbf{w}) \prod_{n=1}^N p(t_n|\mathbf{w})$$

Predictive Distribution

- We may also be interested in making predictions for a new input value \hat{x} .

$$p(\hat{t}|\hat{x}, \mathbf{x}, \mathbf{t}, \alpha, \sigma^2) \propto \int p(\hat{t}, \mathbf{t}, \mathbf{w}|\hat{x}, \mathbf{x}, \alpha, \sigma^2) d\mathbf{w}$$

- where the joint distribution of all the random variables is given by:



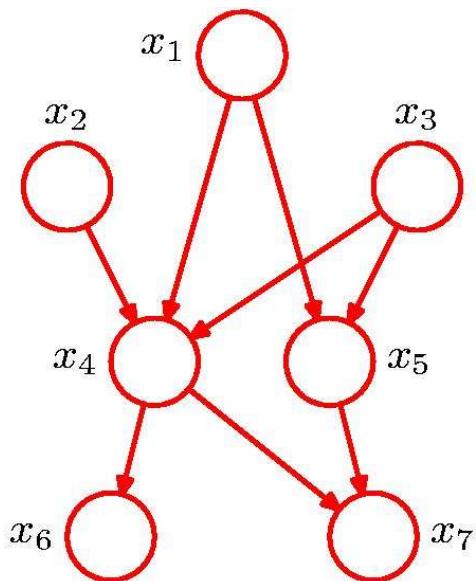
$$p(\hat{t}, \mathbf{t}, \mathbf{w}|\hat{x}, \mathbf{x}, \alpha, \sigma^2) = \left[\prod_{n=1}^N p(t_n|x_n, \mathbf{w}, \sigma^2) \right] p(\mathbf{w}|\alpha)p(\hat{t}|\hat{x}, \mathbf{w}, \sigma^2)$$

- Here we are **setting the random variables in \mathbf{t} to the specific values observed in the data.**

Ancestral Sampling

- Consider a joint distribution over K random variables $p(x_1, x_2, \dots, x_K)$ that factorizes as:

$$p(\mathbf{x}) = \prod_{k=1}^K p(x_k | \text{pa}_k)$$



- Our goal is draw a **sample from this distribution**.
- Start at the top and sample in order.

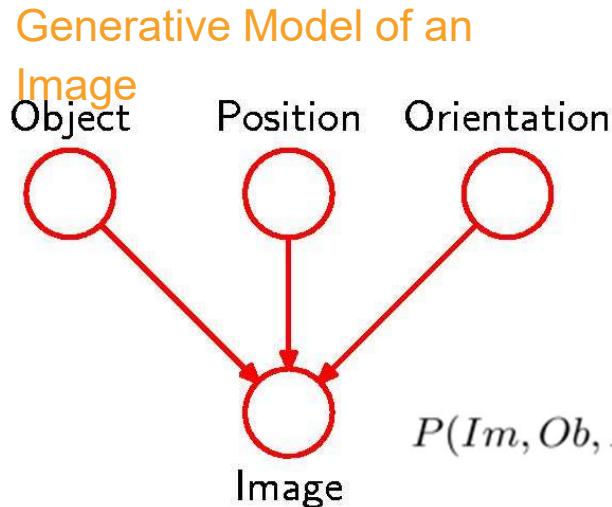
$$\begin{aligned}\hat{x}_1 &\sim p(x_1) \\ \hat{x}_2 &\sim p(x_2) \\ \hat{x}_3 &\sim p(x_3) \\ \hat{x}_4 &\sim p(x_4 | \hat{x}_1, \hat{x}_2, \hat{x}_3) \\ \hat{x}_5 &\sim p(x_5 | \hat{x}_1, \hat{x}_3)\end{aligned}$$

The parent variables are set to their sampled values

- To obtain a sample from **the marginal distribution**, e.g. $p(x_2, x_5)$, we sample from the full joint distribution, retain \hat{x}_2, \hat{x}_5 , and discard the remaining values.

Generative Models

- Higher-level nodes will typically represent **latent (hidden) random variables**.
- The primary role of the latent variables is to allow a complicated distribution over observed variables to be constructed from simpler (**typically exponential family**) conditional distributions.



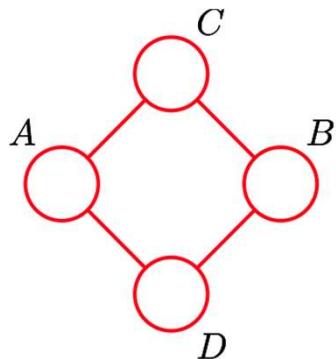
- Object identity, position, and orientation have independent prior probabilities.
- The image has a probability distribution that depends on the object identity, position, and orientation (**likelihood function**).

$$P(Im, Ob, Po, Or) = \underbrace{P(Im|Ob, Po, Or)}_{\text{Likelihood}} \underbrace{P(Ob)P(Po)P(Or)}_{\text{Prior}}$$

- The graphical model captures the **causal process**, by which the observed data was generated (hence the name **generative models**).

Undirected Graphical Models

Directed graphs are useful for expressing causal relationships between random variables, whereas undirected graphs are useful for expressing soft constraints between random variables



- The joint distribution defined by the graph is given by the product of non-negative potential functions over the maximal cliques (connected subset of nodes).

$$p(\mathbf{x}) = \frac{1}{Z} \prod_C \phi_C(x_C) \quad Z = \sum_{\mathbf{x}} \prod_C \phi_C(x_C)$$

where the normalizing constant Z is called a partition function.

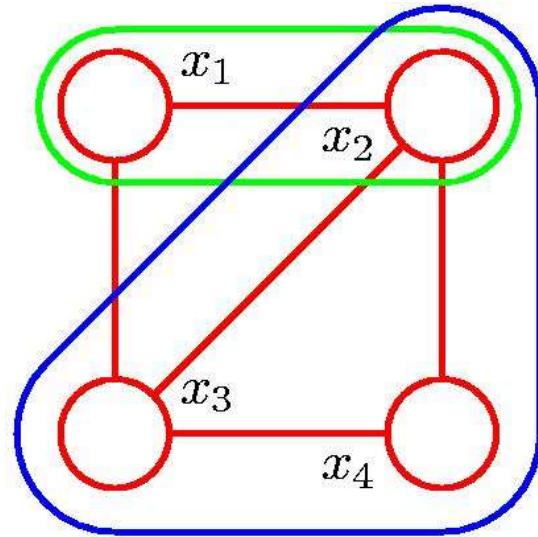
- For example, the joint distribution factorizes:

$$p(A, B, C, D) = \frac{1}{Z} \phi(A, C) \phi(C, B) \phi(B, D) \phi(A, D)$$

- Let us look at the definition of cliques.

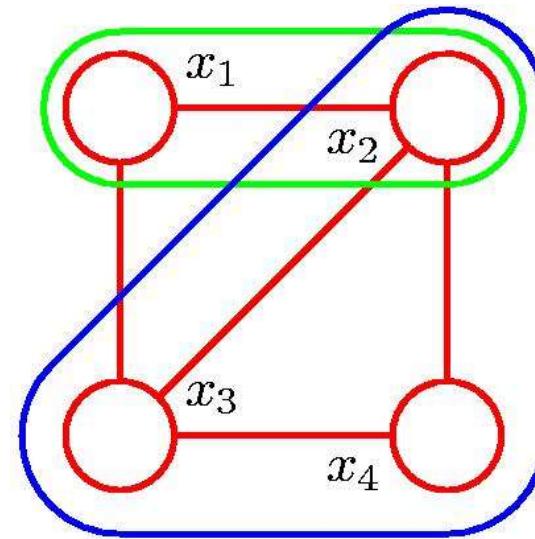
Cliques

- The subsets that are used to define the potential functions are represented by **maximal cliques** in the undirected graph.
- **Clique**: a subset of nodes such that there exists a link between all pairs of nodes in a subset.
- **Maximal Clique**: a clique such that it is not possible to include any other nodes in the set without it ceasing to be a clique.
- This graph has 5 cliques:
 $\{x_1, x_2\}, \{x_2, x_3\}, \{x_3, x_4\},$
 $\{x_4, x_2\}, \{x_1, x_3\}.$
- Two maximal cliques:
 $\{x_1, x_2, x_3\}, \{x_2, x_3, x_4\}.$

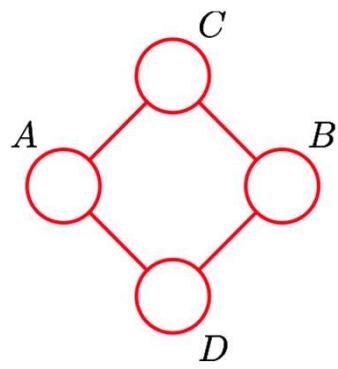


Using Cliques to Represent Subsets

- If the potential functions only involve two nodes, an undirected graph has a nice representation.
- If the potential functions involve more than two nodes, using a different **factor graph representation** is much more useful.
- For now, let us consider only potential functions that are defined over two nodes.



Markov Random Fields (MRFs)



$$p(\mathbf{x}) = \frac{1}{Z} \prod_C \phi_C(x_C)$$

- Each potential function is a mapping from the joint configurations of random variables in a clique to non-negative real numbers.
- The choice of potential functions is not restricted to having specific probabilistic interpretations.

Potential functions are often represented as exponentials:

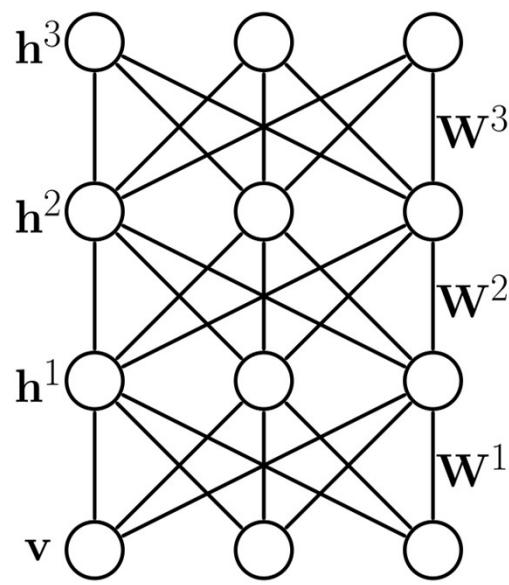
$$p(\mathbf{x}) = \frac{1}{Z} \prod_C \phi_C(x_C) = \frac{1}{Z} \exp\left(-\sum_C E(x_c)\right) = \underbrace{\frac{1}{Z} \exp(-E(\mathbf{x}))}_{\text{Boltzmann distribution}}$$

where $E(\mathbf{x})$ is called an energy function.

Boltzmann
distribution

MRFs with Hidden Variables

For many interesting real-world problems, we need to introduce hidden or latent variables.



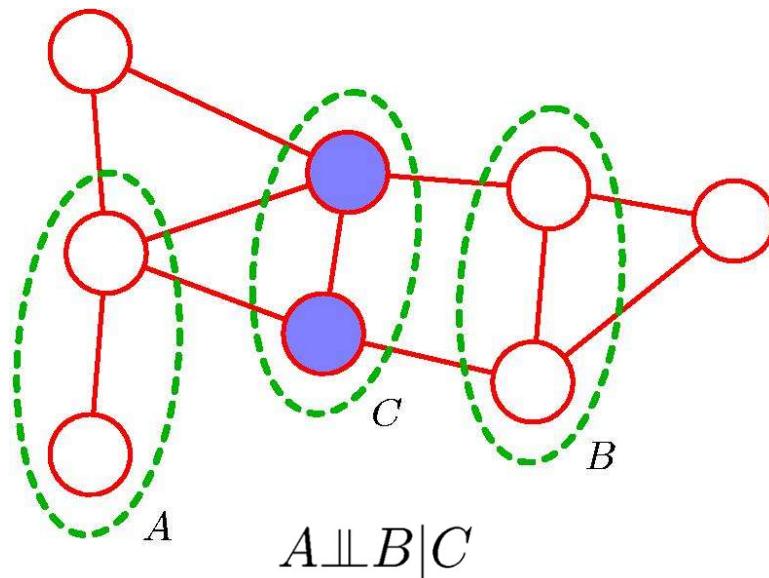
- Our random variables will contain both **visible and hidden** variables $x=(v,h)$.

$$p(\mathbf{v}) = \frac{1}{Z} \sum_{\mathbf{h}} \exp(-E(\mathbf{v}, \mathbf{h}))$$

- In general, computing both partition function and summation over hidden variables will be intractable, except for special cases.
- Parameter learning becomes a very challenging task.

Conditional Independence

- Conditional Independence is easier compared to directed models:



A is conditionally independent of B given C

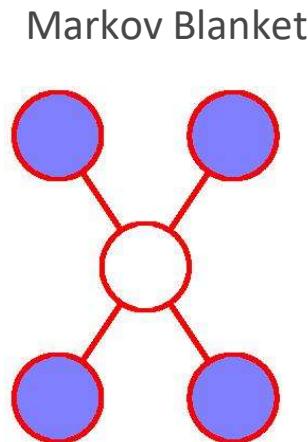
- Observation blocks a node.
- Two sets of nodes are conditionally independent if the observations block all paths between them.

iff every path between A and B is blocked by nodes in C.

A node becomes a “blocker” when it is observed.

Markov Blanket

- The **Markov blanket** of a node is simply all the directly connected nodes.

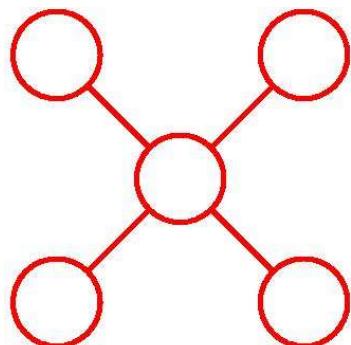


- This is simpler than in directed models, since there is **no explaining away**.
- The conditional distribution of x_i conditioned on all the variables in the graph is dependent only on the variables in the Markov blanket.

Markov blanket in MRF = the set of neighboring nodes

Conditional Independence and Factorization

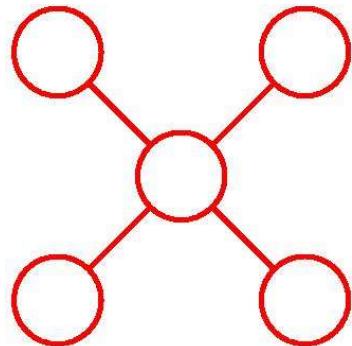
- Consider two sets of distributions:
 - The set of distributions consistent with the conditional independence relationships defined by the undirected graph.
 - The set of distributions consistent with the factorization defined by potential functions on maximal cliques of the graph.
- The **Hammersley-Clifford theorem** states that these two sets of distributions are the same.



$$p(\mathbf{x}) = \frac{1}{Z} \prod_C \phi_C(x_C)$$

Interpreting Potentials

- In contrast to directed graphs, the potential functions **do not have a specific probabilistic interpretation.**



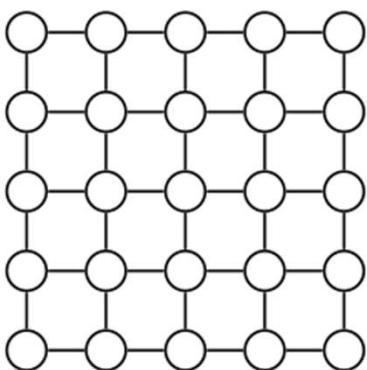
$$p(\mathbf{x}) = \frac{1}{Z} \prod_C \phi_C(x_C) = \frac{1}{Z} \exp\left(-\sum_C E(x_c)\right)$$

- This gives us greater flexibility in choosing the potential functions.

- We can view the potential function as expressing which configuration of the **local variables** are preferred to others.
- Global configurations with relatively high probabilities are those that find a good balance in satisfying the (possibly conflicting) influences of the clique potentials.
- So far we did not specify the nature of random variables, discrete or continuous.

Discrete MRFs

- MRFs with all discrete variables are widely used in many applications.
- MRFs with **binary variables** are sometimes called **Ising models** in statistical mechanics, and **Boltzmann machines** in machine learning literature.



- Denoting the binary valued variable at node j by $x_j \in \{0, 1\}$, the Ising model for a graph $G(V, E)$ and for the joint probabilities is given by:
$$P_\theta(\mathbf{x}) = \frac{1}{Z(\theta)} \exp \left(\sum_{ij \in E} x_i x_j \theta_{ij} + \sum_{i \in V} x_i \theta_i \right)$$
- The conditional distribution is given by logistic:

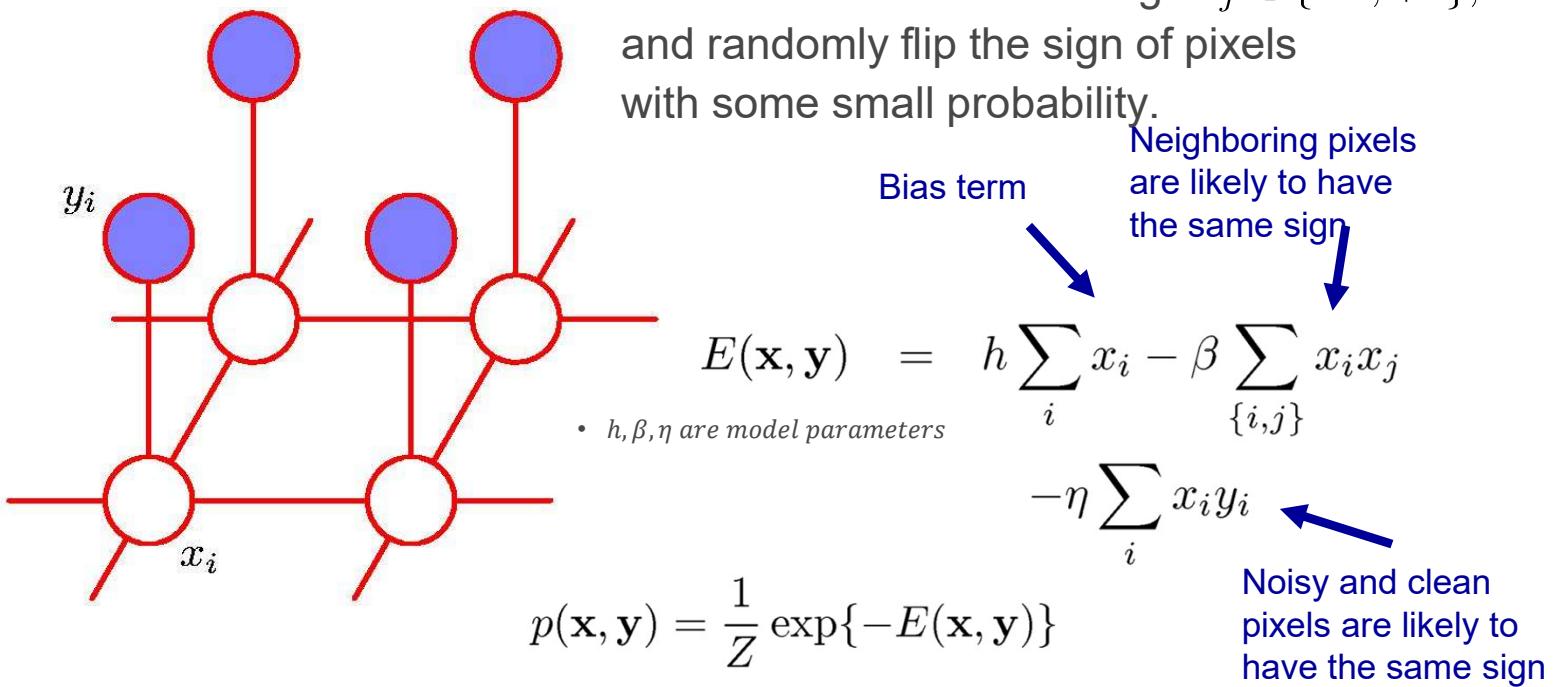
$$P_\theta(x_i = 1 | \mathbf{x}_{-i}) = \frac{1}{1 + \exp(-\theta_i - \sum_{j \in N(i)} x_j \theta_{ij})}, \quad \text{where } \mathbf{x}_{-i} \text{ denotes all nodes except for } i.$$

Hence the parameter θ_{ij} measures the dependence of x_i on x_j , conditional on the other nodes.

Example: Image Denoising

- Let us look at the example of noise removal from a binary image.
- Let the observed noisy image be described by an array of binary pixel values: $y_j \in \{-1, +1\}$, $i=1, \dots, D$.

- We take a noise-free image $x_j \in \{-1, +1\}$, and randomly flip the sign of pixels with some small probability.

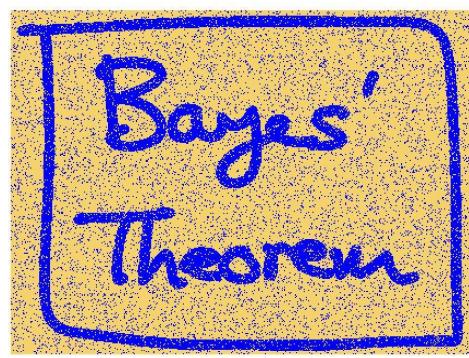


Iterated Conditional Modes

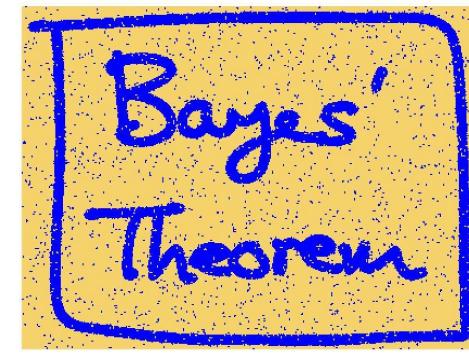
- Iterated conditional modes: coordinate-wise gradient descent.
- Visit the unobserved nodes sequentially and set each x to whichever of its two values has the lowest energy.
 - This only requires us to look at the Markov blanket, i.e. the connected nodes.
 - Markov blanket of a node is simply all the directly connected nodes.



Original
Image



Noisy Image



ICM