

Inference and Representation

DS-GA-1005, CSCI-GA.2569

Joan Bruna

Courant Institute of Mathematical Sciences
Center for Data Science
NYU

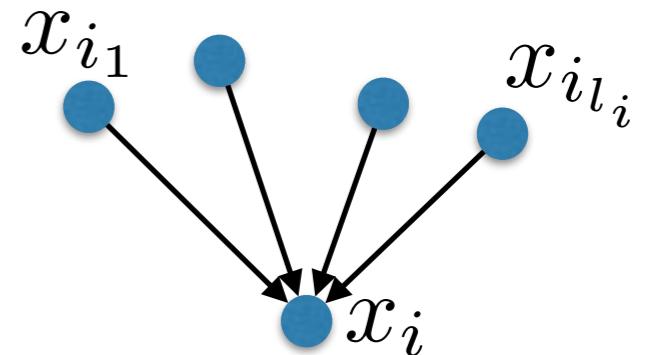


Lecture 1: Bayesian Networks

- We've seen that reducing the scope of conditional probability is critical in order to beat the curse of dimensionality.

$$p_G(x_1, \dots, x_n) = \prod_{i=1}^n p(x_i \mid x_{i_1}, \dots, x_{i_{l_i}}).$$

- Q: How to describe those dependencies?
- Given a directed acyclic graph $G = (V, E)$ we encode a model as
 - One node $i \in V$ for each random variable.
 - One conditional probability distribution per node, where conditional factors become the parents in the graph:
- Properties of the joint distribution are now expressed in terms of the graph.
 - algorithms for inference
 - complexity questions.



$$\{x_{i_1}, \dots, x_{i_{l_i}}\} := x_{Pa}(i)$$

Lect. 1: Conditional Independence

- In a specific graphical model, conditional probabilities have missing terms:

$$p(X_i \mid X_{i_1}, \dots, X_{i_{l_i}}) \text{ instead of } p(X_i \mid X_1, \dots, X_{i-1})$$

- Thus

$$X_i \perp X_j \mid X_{Pa(i)} \text{ for } j < i, j \notin Pa(i).$$

- In fact, missing variables in the local conditional probabilities



missing edges in the corresponding graph

Lecture 1: Joint Distribution – BN

- A Bayesian Network corresponds to a particular factorization of the joint distribution:

$$p(X_1, \dots, X_n) = \prod_{i \in V} p(X_i \mid X_{Pa(i)})$$

- How well a BN corresponds to a given probabilistic model?
 - Given a directed acyclic graph \mathbf{G} , we denote by $I(\mathbf{G})$ the set of conditional independencies implied by it.
 - Given a joint distribution \mathbf{p} , we denote by $I(\mathbf{p})$ the set of conditional independencies satisfied by it.
 - \mathbf{G} is a *perfect independence map* for \mathbf{p} if $I(\mathbf{G}) = I(\mathbf{p})$.
- No unicity: Different BNs can result in the same conditional independencies (different topological order)
- No existence: There exist \mathbf{p} without perfect independence map.

Lecture 2

- Undirected Graphical Models (Markov Random Fields)
 - Definitions
 - Conditional Independence
 - Examples: The Ising Model
- Factor Graphs
- From Bayesian to Markov networks
- Parameter Estimation

Pros/Cons of Directed Models

$$p(X_1, \dots, X_n) = \prod_{i \in V} p(X_i \mid X_{Pa(i)})$$

- Pros:

- They allow efficient sampling (generative model): follow topological order.
- Local factors directly interpreted as conditional probabilities.

- Cons:

- Conditional Independencies are not immediately obvious from the graph (requires notion of d-separation).
- Lack of unicity: different graphs model the same distribution.
- Lack of existence: some distributions do not admit a BN model.

- Q: How to palliate some of its problems?

Undirected Graphical Models

- Consider a graph $G = (V, E)$, with
 - V in 1-to-1 correspondence with our set of rvs X_1, \dots, X_n .
 - E : set of undirected edges within V .

- In the directed case, we started with the parametrization

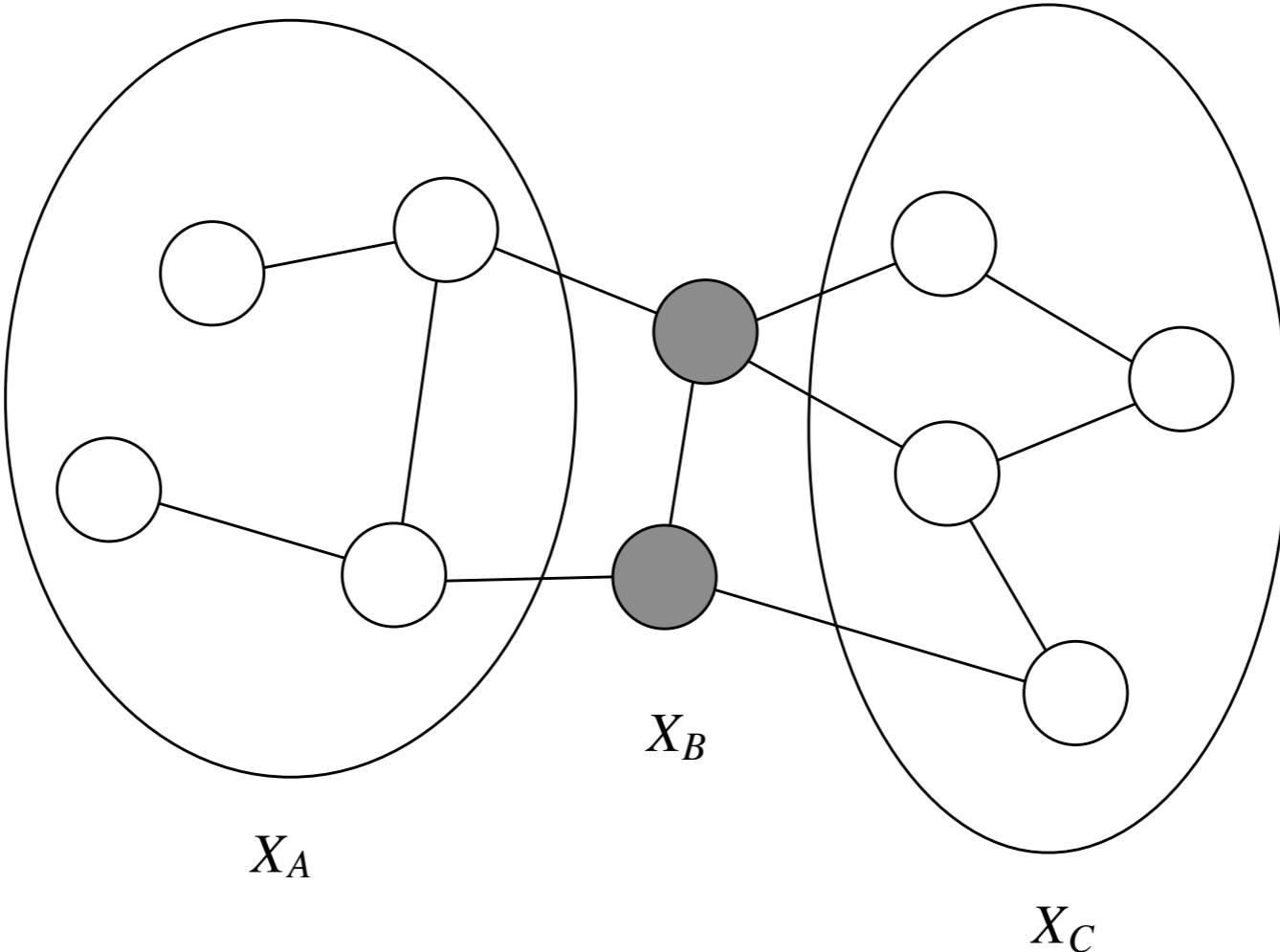
$$p(X_1, \dots, X_n) = \prod_{i \in V} p(X_i \mid X_{Pa(i)})$$

- and then derived the set of conditional independencies.
- In the undirected case, we do the opposite.

Given subsets of nodes A , B and C ,

$X_A \perp\!\!\!\perp X_B \mid X_C$ when nodes X_B separate nodes X_A from X_C .

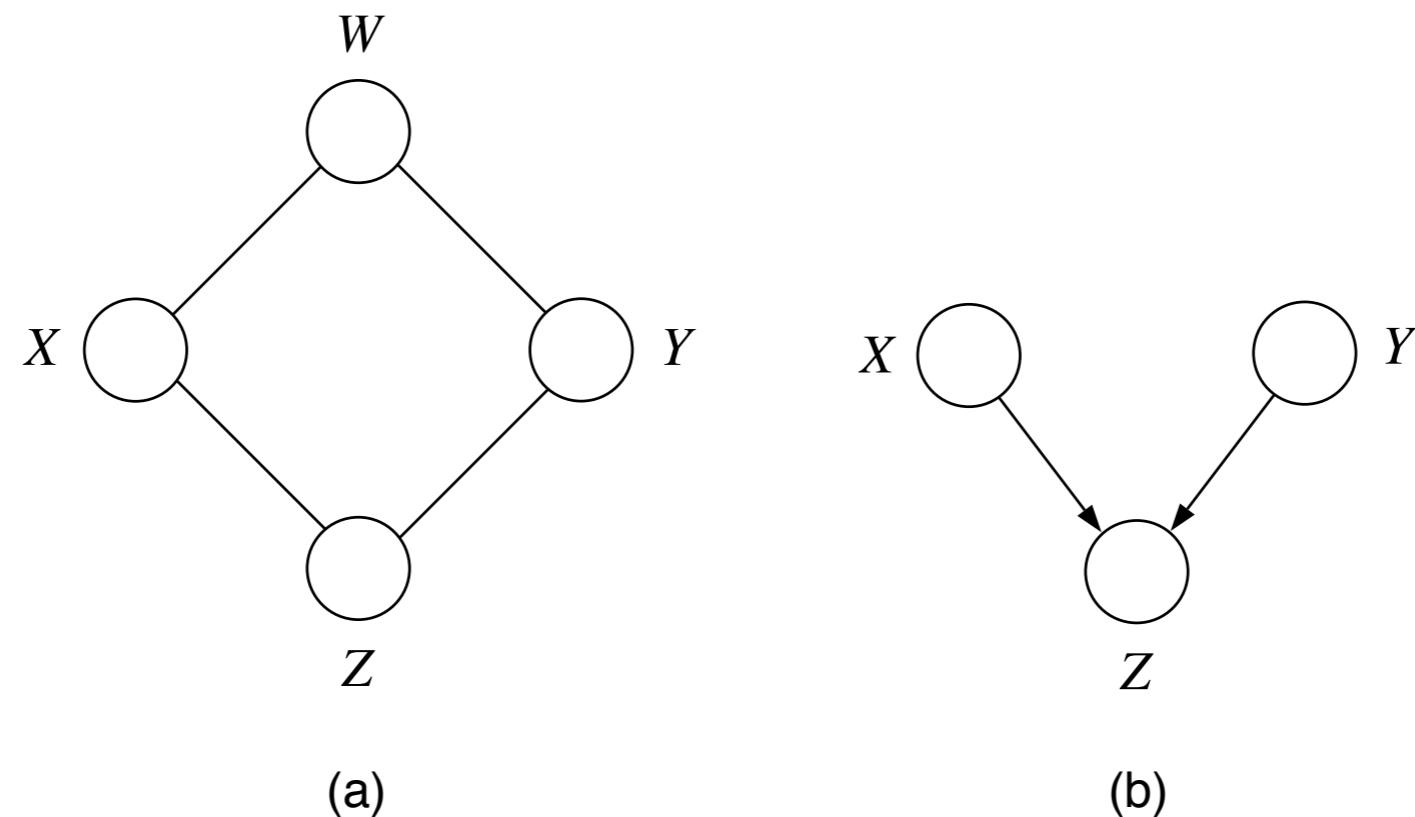
Undirected Graphical Models



- Remark:
 - When we say $X_A \perp X_B \mid X_C$ holds for a given graph \mathbf{G} , we mean that it is true for every probability distribution associated with \mathbf{G}
 - When we say that $X_A \perp X_B \mid X_C$ does not hold for a given \mathbf{G} means that some distributions that factor according to \mathbf{G} do not satisfy it.

Directed vs Undirected

- Q: Can we reduce undirected models to directed ones, and viceversa?



- In (a), we have the conditional independences

$$X \perp Y \mid \{W, Z\}, \quad W \perp Z \mid \{X, Y\} .$$

- ❖ if we want to make it directed, we need to introduce a v-structure!
- In (b), no undirected 3-node graph can represent $X \perp Y$
(without implying also $X \perp Z$ or $Y \perp Z$)

Undirected Graphical Models

- Q: How to parametrize those models? (from graph to distribution)
- In the directed setting, we used conditional probabilities as “building blocks” for the joint distribution:
 - properly normalized thanks to the Bayes chain rule.
- We cannot do this now: how to select “parents” from children locally so that it is globally consistent?
- We still want to obtain a factorization of p as a product of “local” factors.

Undirected Graphical Models

- We want to obtain a factorization $p(x_1, \dots, x_n) = \prod_{C \in \mathcal{C}} \psi_C(x_C)$
 - for appropriate factors $\{C, \psi_C\}_{C \in \mathcal{C}}$
- Consider a pair of nodes X_i and X_j not neighbors in G .
 - It follows that $X_i \perp X_j \mid \{X_k\}_{k \neq i, j}$.
 - Hence

$$\begin{aligned} p(x_i, x_j \mid \{x_k\}_{k \neq i, j}) &= \frac{p(x_1, \dots, x_n)}{p(\{x_k\})_{k \neq i, j}} \\ &= \frac{\prod_{C; i \in C, j \in C} \psi_C(x_C) \cdot \prod_{C; i \in C, j \notin C} \psi_C(x_C) \cdot \prod_{C; i \notin C, j \in C} \psi_C(x_C) \cdot \prod_{C; i \notin C, j \notin C} \psi_C(x_C)}{\int dx_i dx_j p(x_1, \dots, x_n)} \\ &= \frac{1}{Z} \prod_{C; i \in C, j \in C} \psi_C(x_C) \cdot \prod_{C; i \in C, j \notin C} \psi_C(x_C) \cdot \prod_{C; i \notin C, j \in C} \psi_C(x_C) \end{aligned}$$

- Suppose one of the factors C contains both x_i and x_j
- In general, this will violate

$$p(x_i, x_j \mid \{x_k\}) = p(x_i \mid \{x_k\}_k) \cdot p(x_j \mid \{x_k\}_k)$$

Undirected Graphical Models

- Thus, factors only contain nodes that are fully-connected — this is called a *clique*.
- Since a clique of size m contains all cliques of smaller sizes, we can reduce ourselves to *maximal cliques* (cliques that cannot be extended while being fully connected).
 - If X_C form a maximal clique, arbitrary functions $\psi(x_C)$ capture all possible dependencies within the clique.
- So, by considering

\mathcal{C} = set of maximal cliques of G

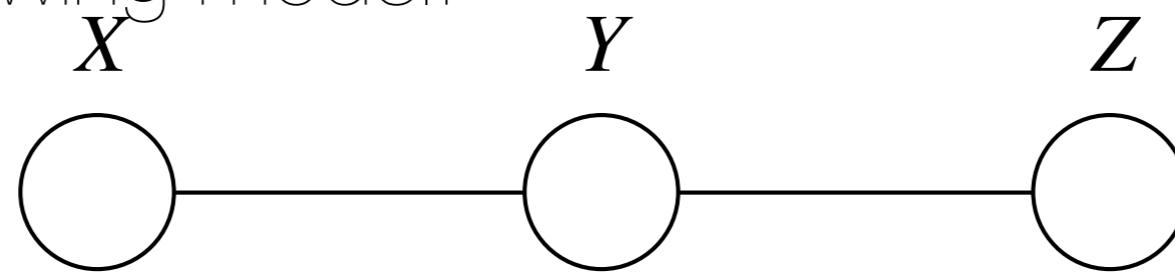
$\psi_C(x_C)$: non-negative potential function (not necessarily normalized)

- We have $p(x) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(x_C)$, $Z = \int dx \prod_{C \in \mathcal{C}} \psi_C(x_C)$.
partition function

Interpretation of Potential Functions

- What is the meaning of the clique-potential functions ψ_C ?
 - not conditional probabilities
 - how about marginal probabilities?

- Consider the following model:



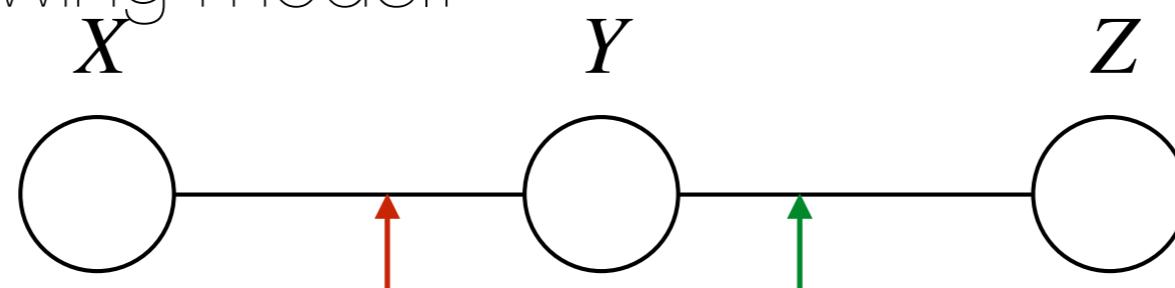
We have $X \perp Z \mid Y$.

Thus $p(x, y, z) = p(x \mid y)p(y)p(z \mid y)$

Interpretation of Potential Functions

- What is the meaning of the clique-potential functions ψ_C ?
 - not conditional probabilities
 - how about marginal probabilities?

- Consider the following model:



We have $X \perp Z \mid Y$.

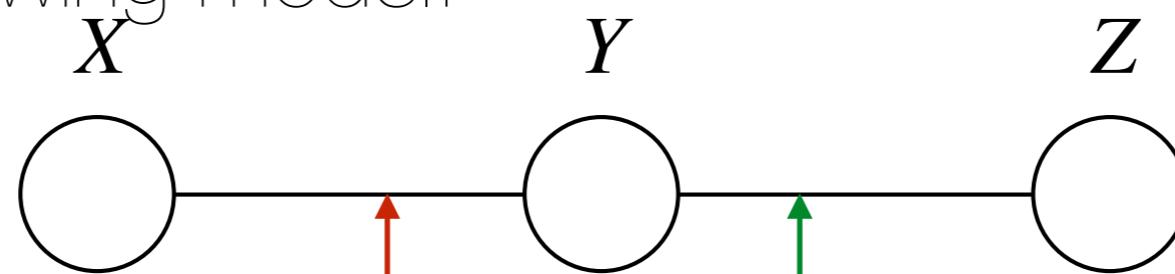
Thus $p(x, y, z) = p(x \mid y)p(y)p(z \mid y)$

marginal conditional

Interpretation of Potential Functions

- What is the meaning of the clique-potential functions ψ_C ?
 - not conditional probabilities
 - how about marginal probabilities?

- Consider the following model:



We have $X \perp Z \mid Y$.

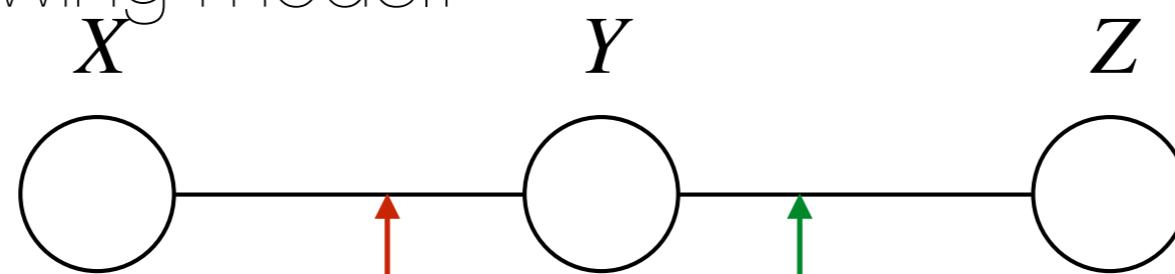
Thus $p(x, y, z) = p(x \mid y)p(y)p(z \mid y)$

conditional marginal

Interpretation of Potential Functions

- What is the meaning of the clique-potential functions ψ_C ?
 - not conditional probabilities
 - how about marginal probabilities?

- Consider the following model:



We have $X \perp Z \mid Y$.

Thus $p(x, y, z) = p(x \mid y)p(y)p(z \mid y)$

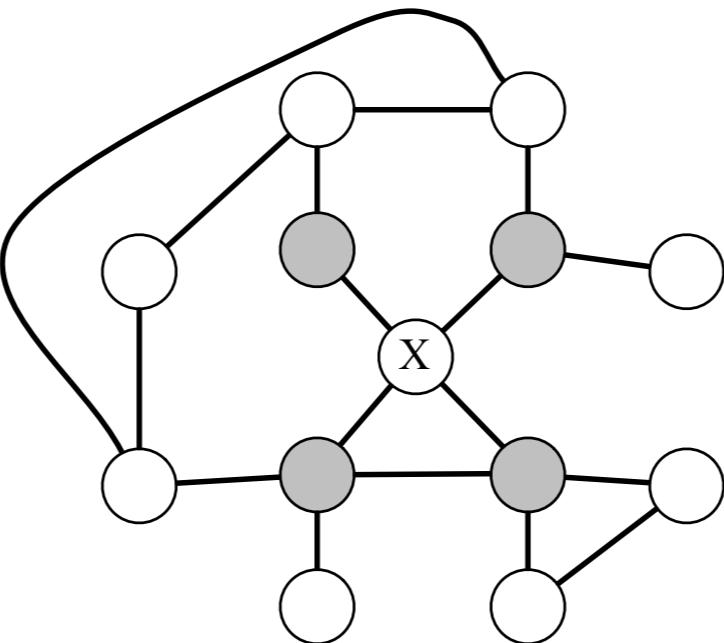
conditional marginal

- In general, $p(x, y, z) \neq p(x, y)p(y, z)$.

$$p(x, y, z) = p(x, y)p(y, z) \Rightarrow p(y) = 0 \text{ , or } p(y) = 1 .$$

Markov Blanket

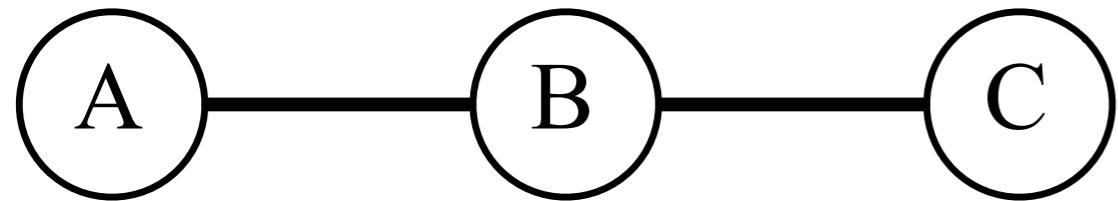
- A set $A \subseteq \mathcal{X}$ is a Markov Blanket of X if $X \notin A$ and if A is a minimal set of nodes such that $X \perp (\mathcal{X} \setminus (A \cup X)) \mid A$.
- In undirected graphical models, the Markov Blanket of a variable is precisely its neighbors in the graph:



- X is independent of the rest of nodes conditioned on its neighbors.

Independence Through Separation

- We illustrate notion of Markov Blanket with the following example:



$$p(a, b, c) = \frac{1}{Z} \psi_{AB}(a, b) \psi_{BC}(b, c) .$$

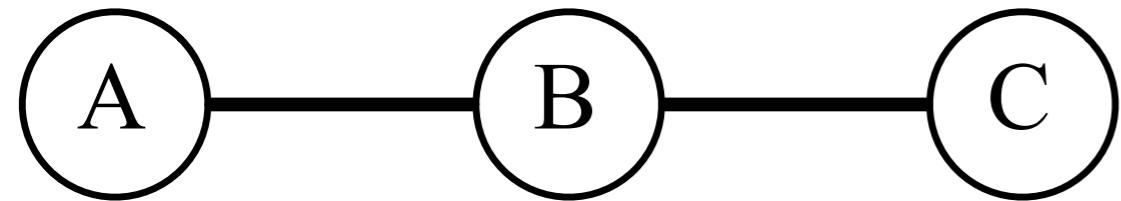
- Let us verify $A \perp C \mid B$

$$\begin{aligned} p(a \mid b) &= \frac{p(a, b)}{p(b)} = \frac{\sum_{c'} \psi_{AB}(a, b) \psi_{BC}(b, c')}{\sum_{a', c'} \psi_{AB}(a', b) \psi_{BC}(b, c')} \\ &= \frac{(\sum_{c'} \psi_{BC}(b, c')) \psi_{AB}(a, b)}{(\sum_{c'} \psi_{BC}(b, c')) (\sum_{a'} \psi_{AB}(a', b))} \\ &= \frac{\psi_{AB}(a, b)}{\sum_{a'} \psi_{AB}(a', b)} \end{aligned}$$

- In general, separability holds when a variable is conditioned on its Markov Blanket.

Independence Through Separation

- We illustrate notion of Markov Blanket with the following example:



$$p(a, b, c) = \frac{1}{Z} \psi_{AB}(a, b) \psi_{BC}(b, c) .$$

- It follows that

$$\begin{aligned} p(a, c \mid b) &= \frac{p(a, b, c)}{p(b)} = \frac{\psi_{AB}(a, b) \psi_{BC}(b, c)}{\sum_{a', c'} \psi_{AB}(a', b) \psi_{BC}(b, c')} \\ &= \frac{\psi_{AB}(a, b)}{\sum_{a'} \psi_{AB}(a', b)} \cdot \frac{\psi_{BC}(b, c)}{\sum_{c'} \psi_{BC}(b, c')} \\ &= p(a \mid b) p(c \mid b) . \end{aligned}$$

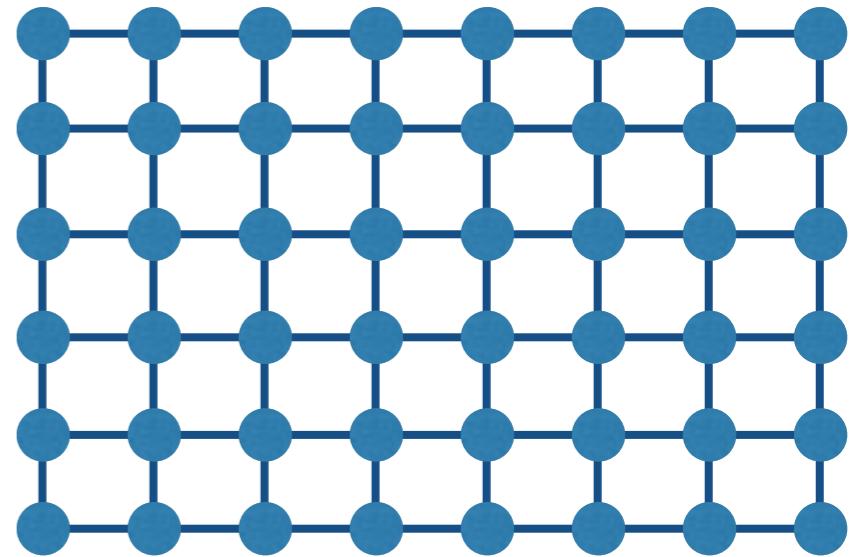
- Also, conditional probabilities do not “care” about partition function.

Example: Ising Model

- Consider a collection of *spins* or magnets, taking discrete ± 1 values, arranged in a lattice configuration.
- Spins interact with their neighbors in the lattice, either by attracting each other to have same state (ferromagnetism) or by repulsing each other (anti-ferromagnetism).



1d lattice (Ising, 1924)



2d lattice (Onsager, 1944)

$i \in \mathcal{L}$: 1d/2d Lattice $X_i = \pm 1$: state of each spin

$$p(X_1, \dots, X_n) = \frac{1}{Z} \exp \left(\sum_{i < j} w_{i,j} X_i X_j - \sum_i u_i X_i \right).$$

Ising Model

$$p(X_1, \dots, X_n) = \frac{1}{Z} \exp \left(\sum_{i < j} w_{i,j} X_i X_j - \sum_i u_i X_i \right).$$

- Undirected graphical model with graph given by (1d/2d) lattice.
 - $w_{i,j} > 0$: ferromagnetic interactions (why?)
 - $w_{i,j} < 0$: anti-ferromagnetic interactions (why?)
 - u_i : external magnetic field
 - only neighbors in the lattice contribute to the interaction terms.
- From statistical mechanics, we can interpret the exponent

$$H(X) = - \sum_{i < j} w_{i,j} X_i X_j - \sum_i u_i X_i$$

as an energy quantity (in fact, it is the Hamiltonian of the system).

Boltzmann/Gibbs distributions

- Boltzmann/Gibbs distribution:

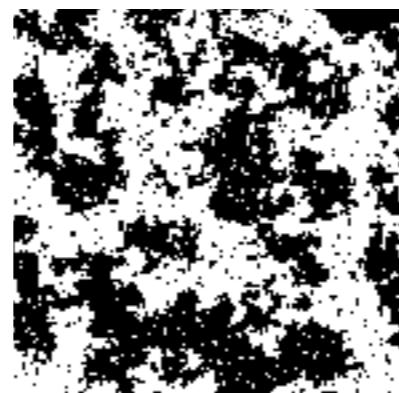
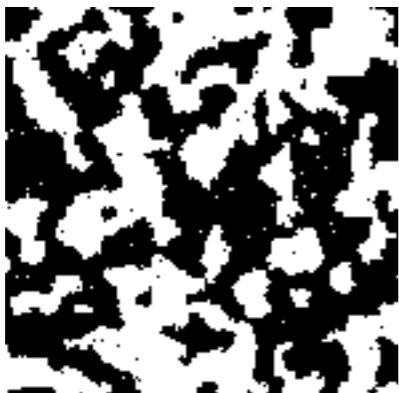
$$p(X) = \frac{1}{Z} \exp(-\beta H(X))$$

- Foundational in the development of statistical mechanics and thermodynamics.

- **β : status of a temperature parameter.**

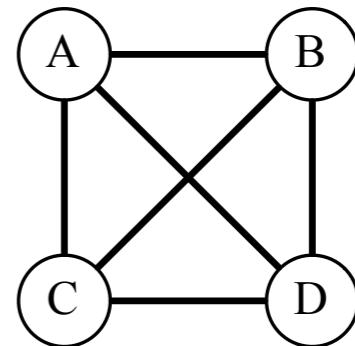
- controls the “order” of the system.
- at high temperatures, particles tend to behave more independently from each other.
- at low temperatures, the system “locks” to specific configurations.
- Phase transitions: study of critical phenomena that goes from “ordered” to “disordered”.

- Ex:



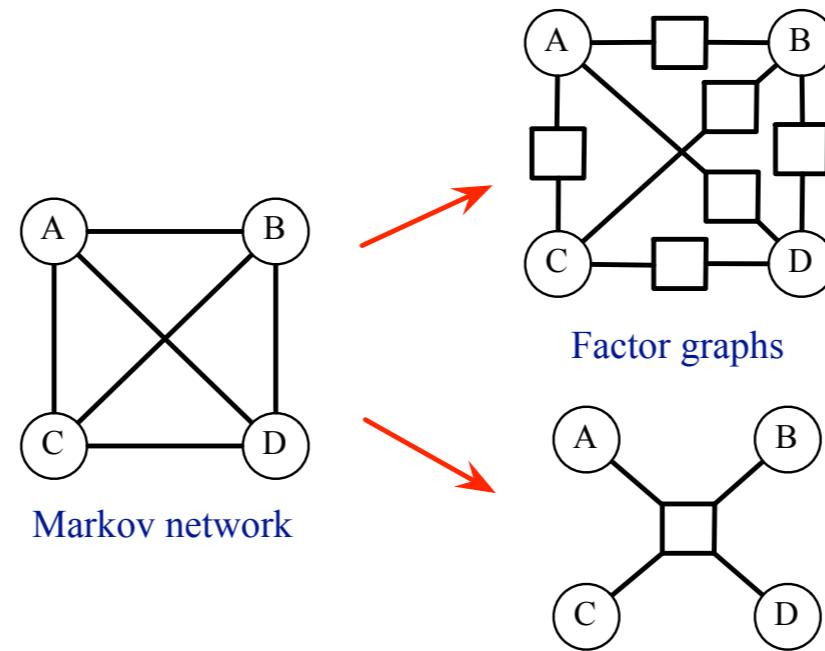
Factor Graphs

- Recall clique vs maximal clique ambiguity.
- Q: Does this have all pair-wise potentials, or a single joint potential of 4 variables?
- Fact: The 4-clique $\psi(a, b, c, d)$ is strictly more general than the pairwise factorization $p(a, b, c, d) = \frac{1}{Z} \prod_{i \neq j} \psi_{ij}(i, j)$
- Q: How to resolve the ambiguity of the graphical representation?



Factor Graphs

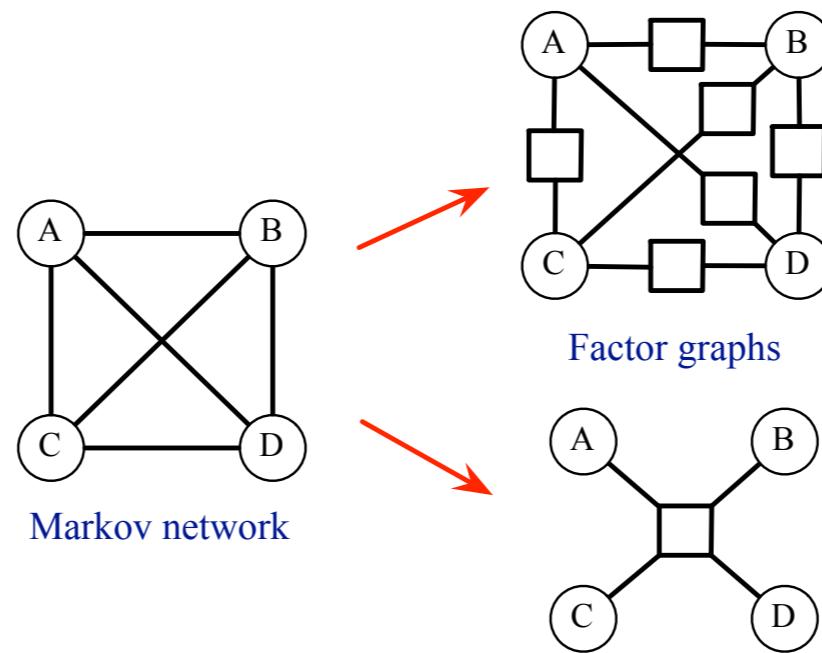
- A factor graph is a bipartite graph where
 - nodes correspond to **both** random variables $\{X_i\}_{i \leq n}$ and potential factors $\{\psi_C\}_{C \in \mathcal{C}}$.
 - edges can only be drawn between variable and factor nodes (if variable X_i appears in factor ψ_C).



- Factor graphs do not have the clique vs maximal clique ambiguity (why?). Same probabilistic model, different graphical representation.

Factor Graphs

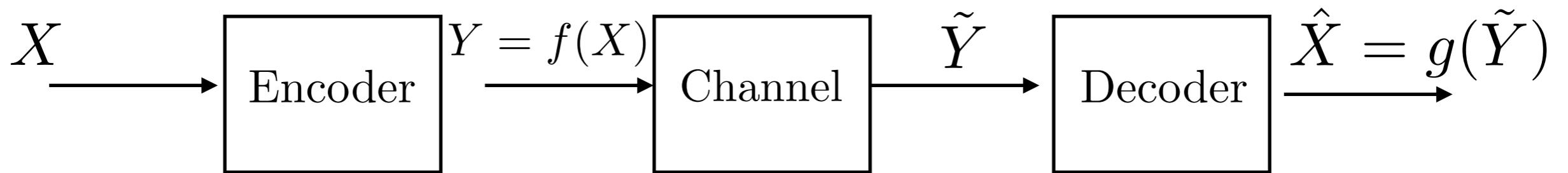
- A factor graph is a bipartite graph where
 - nodes correspond to **both** random variables $\{X_i\}_{i \leq n}$ and potential factors $\{\psi_C\}_{C \in \mathcal{C}}$.
 - edges can only be drawn between variable and factor nodes (if variable X_i appears in factor ψ_C).



- Also, any Bayesian Network or any Markov Random Field can be represented as factor graphs (why?)

Example: LDPC Codes

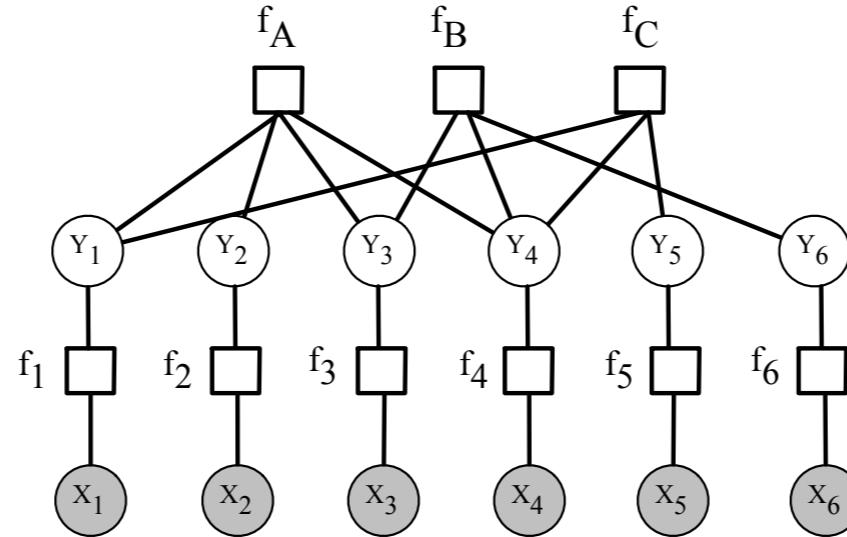
- Error-correcting codes are ubiquitous in digital communications.
- Basic Principle: given a sequence $X = (X_1, \dots, X_n)$ of discrete (e.g binary) to be transmitted, encode it to $Y = f(X)$ of $m \geq n$ bits prior to transmit, in order to increase error robustness.



- Long history in Information Theory
 - Parity codes
 - Also Group Theory and Algebraic Geometry.

LDPC Codes

- Low-density parity-check codes (Gallager, 60s):



- Each of the top row factors provides a parity constraint:

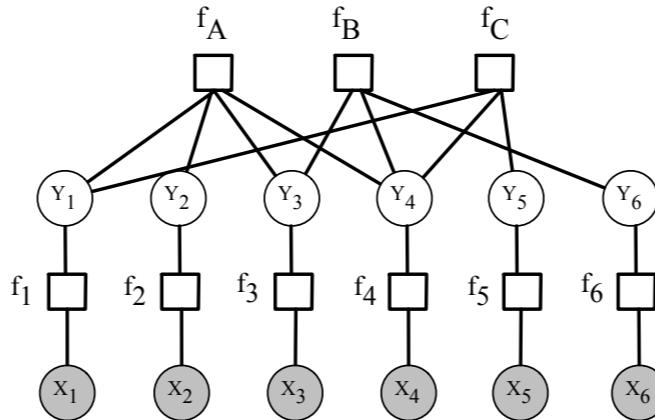
$$f_A(Y_1, Y_2, Y_3, Y_4) = \begin{cases} 1 & \text{if } Y_1 \oplus Y_2 \oplus Y_3 \oplus Y_4 = 0 , \\ 0 & \text{otherwise.} \end{cases}$$

- The three top factors together imply that the only assignments \mathbf{Y} with non-zero probability are:

000000, 011001, 110010, 101011, 111100, 100101, 001110, 010111

- Noise model: $f_i(X_i, Y_i) = p(X_i | Y_i)$.

LDPC Codes



- Q: How to decode from observed data?

$$\hat{Y} = \arg \max_{\mathbf{Y}} p(\mathbf{Y} \mid X = x) \quad (\text{Maximum a Posteriori inference})$$

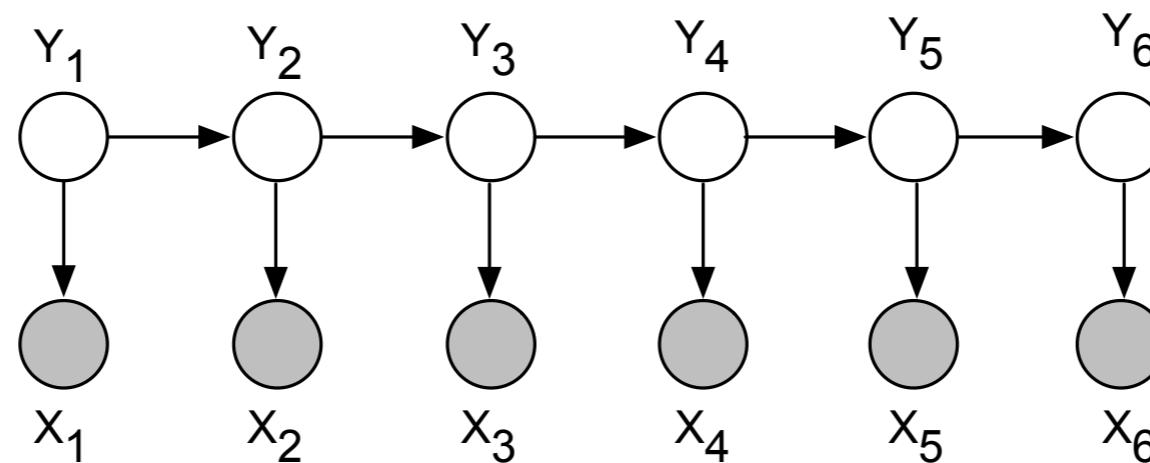
- For general $p(\mathbf{Y}, X) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(Y_C, X_C)$, we have

$$\hat{Y} = \arg \max_{\mathbf{Y}} \sum_{C \in \mathcal{C}} \log \psi_C(Y_C, X_C)$$

- We will see several approaches to solve this later on
 - Belief Propagation on trees.

BNs and MRFs

- So far, we have seen two formalisms to model joint distributions with graphs:
 - Directed Acyclic from Bayes Chain Rule: Bayesian Networks.
 - Undirected graphs: Markov Random Fields.
- Q: To what extent we can “translate” between formalisms?
- Ex: What is the equivalent undirected network for the HMM?



Moralization

- Algorithm to map a Bayesian Network to a Markov Network.
- Given $G = (V, E)$ DAG, we define $M(G)$ an undirected graph over V , with edge between X_i and X_j whenever
 - $X_j \rightarrow X_i$ or $X_i \rightarrow X_j$ in G .
 - X_i and X_j are parents of the same node.



- In $M(G)$, we can no longer tell that $A \perp B$.
 - V-structures disappear, but we can still model "explaining away" with e.g. sparsity priors.

Moralization

- Equivalently, this rule is obtained by mapping factorization of joint distribution.

Bayesian Net

$$p(x_1, \dots, x_n) = \prod_i p(x_i \mid x_{Pa(i)})$$

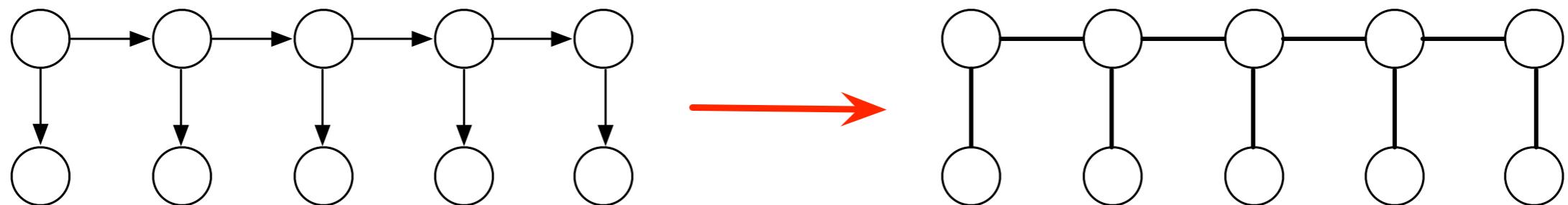
MRF

$$p(x_1, \dots, x_n) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(x_C)$$

- Each node generates a factor in the resulting factor graph:

$$\psi_{C_i}(x_{C_i}) := p(x_i \mid x_{Pa(i)}) , \quad C_i = \{i\} \cup Pa(i) .$$

- Ex: Hidden Markov Model:



Hammersley-Clifford Theorem

- We saw last week that some distributions cannot be modeled as Bayesian Networks.
- Now we ask: which distributions can be written as Markov Fields using an appropriate graph?

- $p(x)$ is a *Gibbs distribution over G* if it can be written as

$$p(x_1, \dots, x_n) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(x_C), \quad \mathcal{C} = \text{cliques of } G$$

- We saw earlier that

If p is a Gibbs distribution for G , then $I(G) \subseteq I(p)$.

– i.e. if Y separates X and Z in G , then $X \perp Z \mid Y$.

- Converse true?

– Not in general.

Hammersley-Clifford Theorem

- However, if we assume that p is positive, i.e. $p(x) > 0$ for all x ,
- Then we have

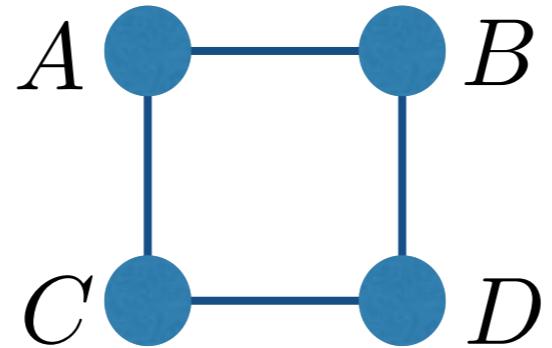
Theorem [H-C]: An undirected graph G is an I-map for a positive distribution $p(x)$ iff p is a Gibbs distribution that factorizes over G .

- It provides a parametrization for any distribution that complies with a series of conditional independence assumptions.
- Positivity condition is needed!

Global Markov but not Factorizing

- Consider 4 binary random variables A, B, C, D, and the following distribution:

$$p(A = 1, B = 1, C = 1, D = 1) = \frac{1}{8}, \quad p(A = 1, B = 1, C = 0, D = 1) = \frac{1}{8}$$
$$p(A = 0, B = 1, C = 0, D = 1) = \frac{1}{8}, \quad p(A = 0, B = 0, C = 0, D = 1) = \frac{1}{8}$$
$$p(A = 0, B = 0, C = 0, D = 0) = \frac{1}{8}, \quad p(A = 0, B = 0, C = 1, D = 0) = \frac{1}{8}$$
$$p(A = 1, B = 0, C = 1, D = 0) = \frac{1}{8}, \quad p(A = 1, B = 1, C = 1, D = 0) = \frac{1}{8}$$



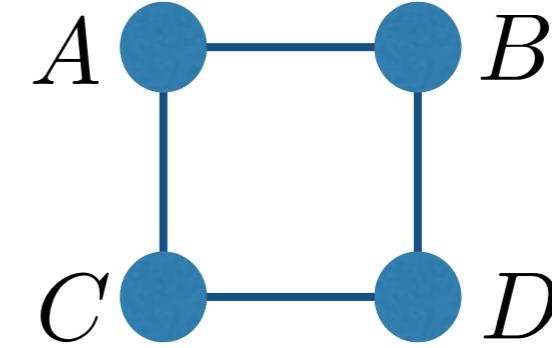
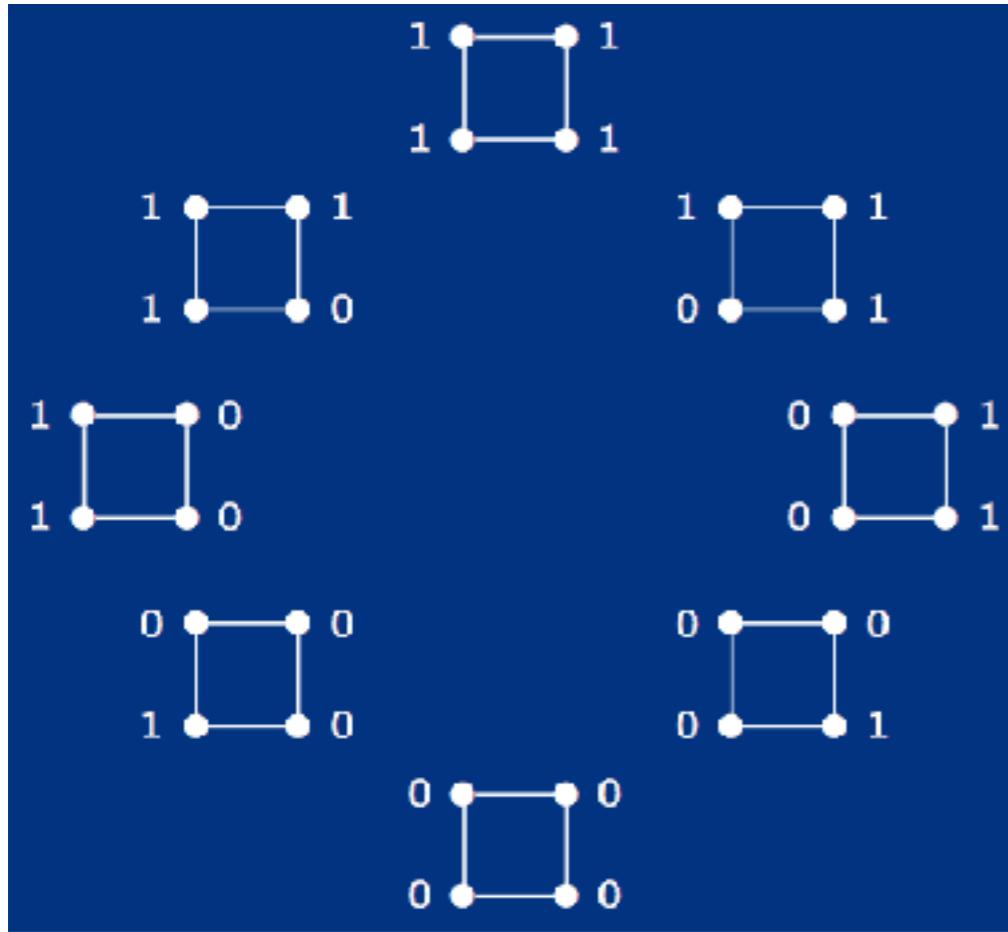
- Do we have $I(G) \subseteq I(p)$?

$$A \perp D \mid \{B, C\} \quad B \perp C \mid \{A, D\}$$

– Observe that conditioning on opposite corners always yields one corner deterministic.

Global Markov but not Factorizing

- Consider 4 binary random variables A, B, C, D, and the following distribution:



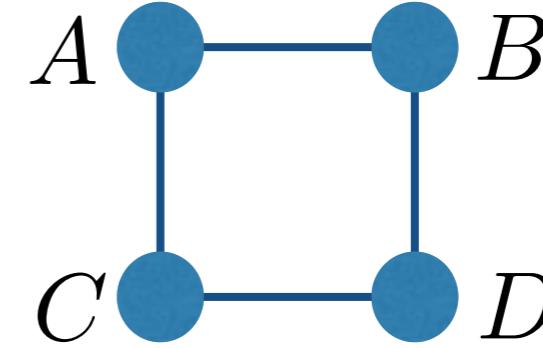
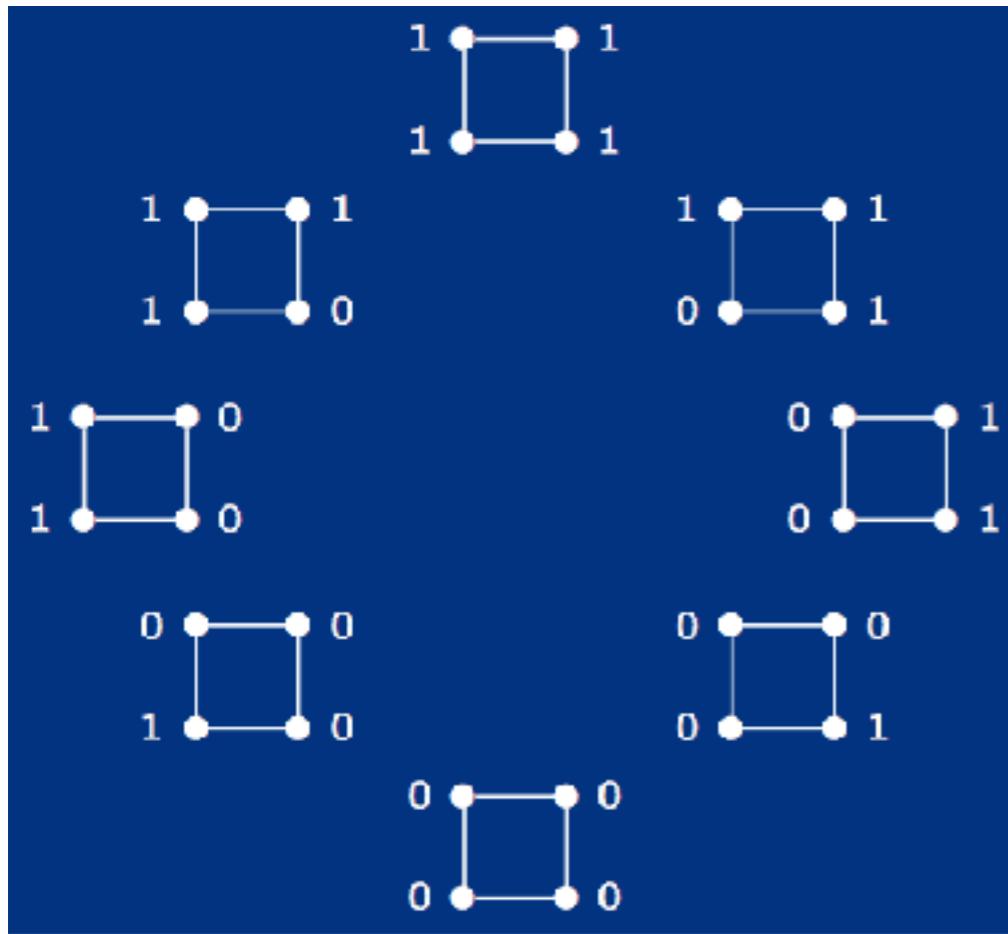
- Do we have $I(G) \subseteq I(p)$?

$$A \perp D \mid \{B, C\} \quad B \perp C \mid \{A, D\}$$

– Observe that conditioning on opposite corners always yields one corner deterministic, and $X \perp Y$ whenever X or Y are deterministic.

Global Markov but not Factorizing

- Consider 4 binary random variables A, B, C, D, and the following distribution:



- Is p a Gibbs distribution?

– Assume $p(x) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(x_C)$

so all these factors are strictly positive

$$0 < Z \cdot p(0, 0, 0, 0) = \psi_{AB}(0, 0)\psi_{BD}(0, 0)\psi_{DC}(0, 0)\psi_{CA}(0, 0)$$

– Trying all 8 positive events implies all factors are strictly positive!

Parameter Estimation

- So far, we have described two families of graphical models, with pros and cons.
- In practice, given some dataset, how to choose which one? Which parameters?
- We assume data is sampled from an underlying (unknown) distribution p^* , associated to some network model $\mathcal{M}^* = (G^*, \theta^*)$
- Samples $\{\mathbf{X}^1, \dots, \mathbf{X}^L\} \sim p^*$ iid.
- In order to "search" for \mathcal{M}^* , we parametrize the search within a family of graphical models
 - We can learn both model parameters for a fixed graph structure,
 - or both structure and parameters.

Task-driven inference

- Depending on the task, we might want to perform different kinds of estimation.
 1. Density Estimation: we are interested in the joint distribution, which can be subsequently used to perform any inference query.
 2. Prediction: we are only interested in a specific set of conditional distribution, e.g classification, or output prediction.
 3. Structural discovery: We are interested in the graph itself (not so much the parameters), e.g. determining dependencies between genes.
- (1) is typically harder than (2). (3) is typically harder than (2).

Parameter Estimation

- Let us focus on (1) first. $\{\mathbf{X}^1, \dots, \mathbf{X}^L\} \sim p^*$ iid.
- Suppose $p^* = p_{\theta^*}$ for some θ^* .
- Two main approaches for parameter estimation:
 - Maximum Likelihood Estimation:

$$E(\theta) = \log p(\{\mathbf{X}^1, \dots, \mathbf{X}^L\} \mid \theta) = \sum_{l \leq L} \log p(\mathbf{X}^l \mid \theta)$$

$$\hat{\theta}_{MLE} = \arg \max_{\theta} E(\theta)$$

- Under appropriate assumptions, $\hat{\theta}_{MLE}$ is
 - ❖ consistent (as sample size grows, $\hat{\theta}_{MLE} \rightarrow \theta^*$ (in probability))
 - ❖ asymptotically efficient (no other consistent estimator has lower asymptotic mean-squared error).
- However, in general this estimation is computationally intractable.

Parameter Estimation

- Let us focus on (1) first. $\{\mathbf{X}^1, \dots, \mathbf{X}^L\} \sim p^*$ iid.
- Suppose $p^* = p_{\theta^*}$ for some θ^* .
- Two main approaches for parameter estimation:
 - Method of Moments:

Consider measurable functions g_1, \dots, g_S .

$$(\text{e.g. } g_i(\mathbf{x}) = x_{i_1}x_{i_2})$$

For each θ , we have $\mu_s(\theta) = \mathbb{E}_{\mathbf{X} \sim p_\theta}(g_s(\mathbf{X}))$ $s = 1 \dots S$

For appropriate choice of moments/functions, system is invertible:

$$\theta = F(\mu)$$

We estimate μ by replacing expectations with empirical averages:

$$\hat{\mu}_s = \frac{1}{L} \sum_{l \leq L} g_s(X^l) \quad s = 1 \dots S$$

And we plug-in the estimator for θ : $\hat{\theta}_{MM} = F(\hat{\mu})$

MLE in Bayesian Networks

- Let us illustrate ML estimation on BN, assuming we know the Bayesian structure \mathbf{G} .

$$p(x_1, \dots, x_n; \theta) = \prod_{i=1}^n p(x_i \mid x_{Pa(i)}; \theta)$$

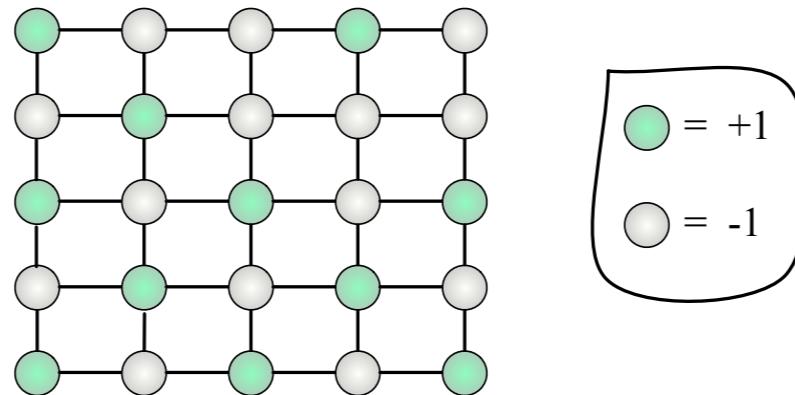
- Given iid samples $\{X^1, \dots, X^L\}$, its log-likelihood is

$$\begin{aligned} E(\theta) &= \sum_{l \leq L} \sum_{i \leq n} \log p(X_i^l \mid X_{Pa(i)}^l; \theta) \\ &= \sum_{i \leq n} \sum_{l \leq L} \log p(X_i^l \mid X_{Pa(i)}^l; \theta_i). \end{aligned}$$

– so the estimation is separable across different factors, breaking the curse of dimensionality.

- Q: How about Markov Random Fields?

Parameter Estimation in MRFs



$$p(x_1, \dots, x_n) = \frac{1}{Z} \exp \left(\sum_{i < j} w_{i,j} x_i x_j - \sum_i u_i x_i \right)$$

- In a MRF, we also have a factorization into local potentials...

$$p(x_1, \dots, x_n; \theta) = \frac{1}{Z(\theta)} \prod_{C \in \mathcal{C}} \psi_C(x_C; \theta) .$$

- ... but the partition function entangles the estimation!

$$\sum_{l \leq L} \log p(X^l; \theta) = \sum_{l \leq L} \left(\sum_{C \in \mathcal{C}} \log \psi(X_C^l; \theta) - \log Z(\theta) \right) .$$

Lecture 2 Summary