NUS National University of Singapore | School of Computing

# CS5340
# Uncertainty Modeling in AI

## Lecture 4:
## Markov Random Fields
## (Undirected Graphical Models)

Asst. Prof. Harold Soh

AY 2022/23

Semester 2

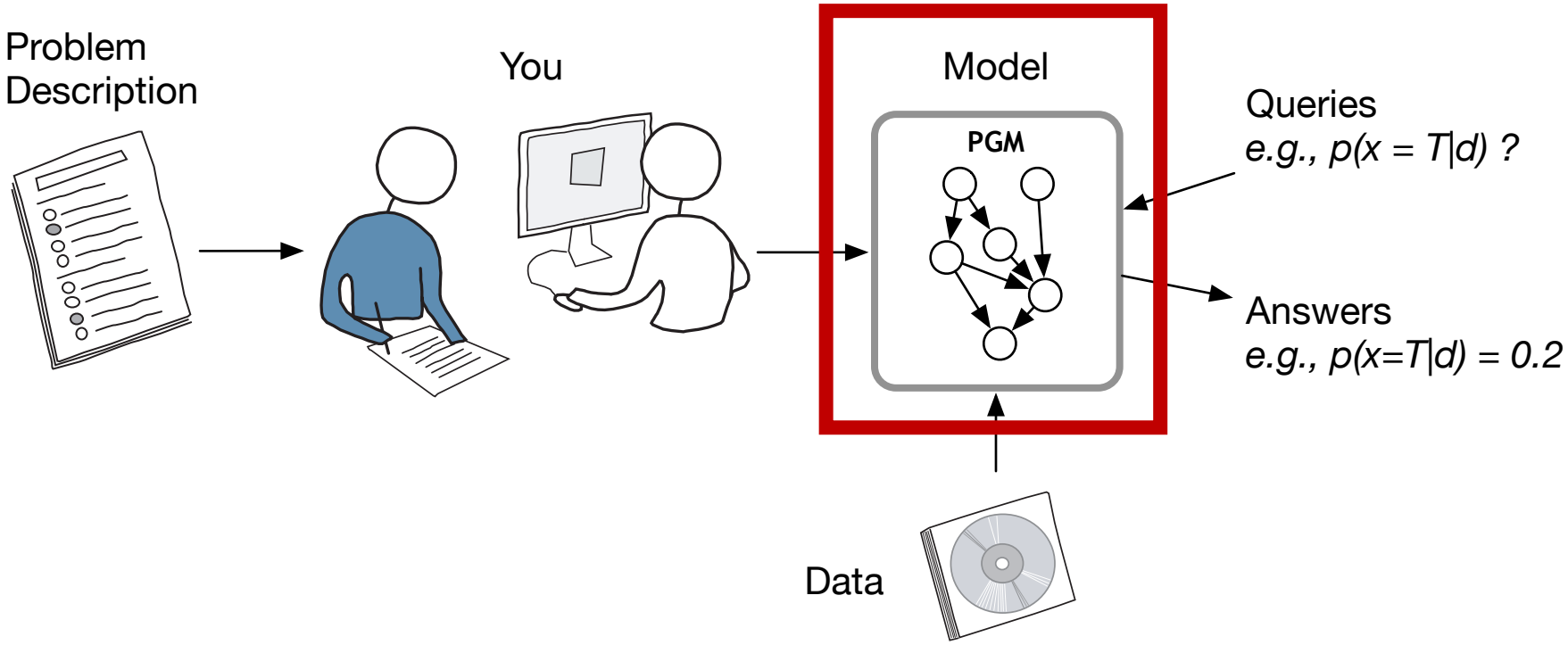
NUS National University of Singapore | School of Computing

# Recap from Lecture 3

*DGM Notation, Examples, and Theoretical Foundations*

# CS5340 in a nutshell

CS5340 is about how to "**represent**" and "**reason**" with uncertainty in a computer.

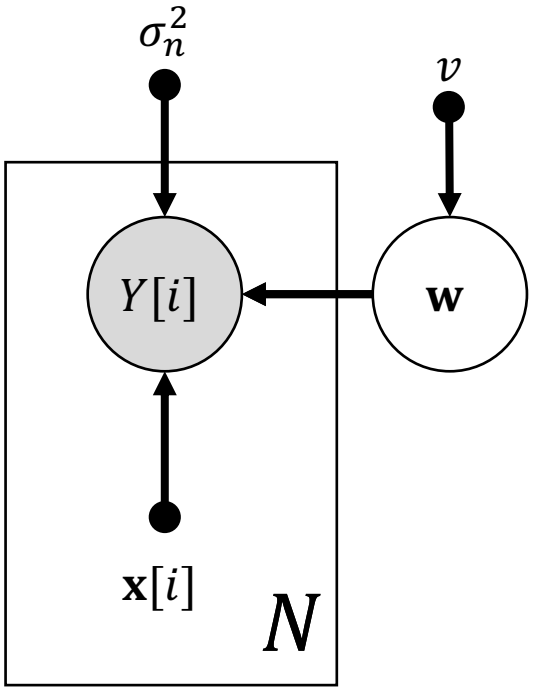


# Bayes Nets (BN) and I-maps

- **Definition** (*Bayesian Network*) A Bayesian network is a tuple $B = (G, P)$ where $P$ factorizes according to $G$ and where $P$ is specified as a set of conditional probability distributions (CPDs) associated with $G$'s nodes.

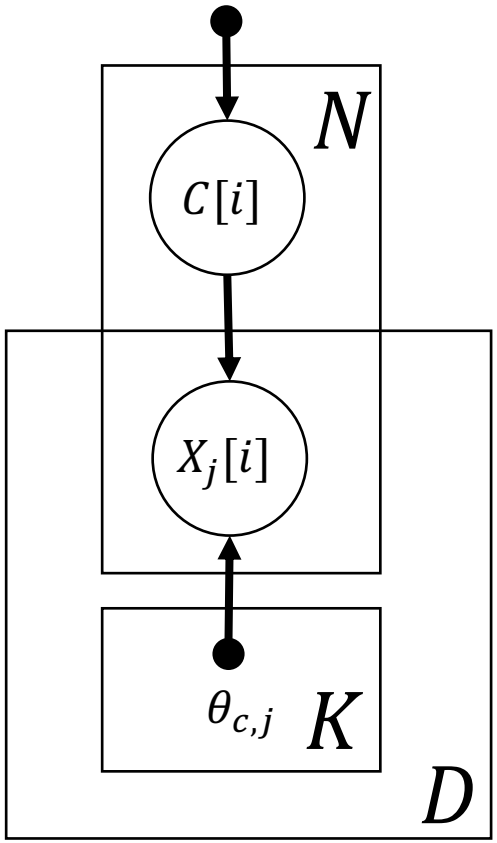
$$p(x_1, \ldots, x_N) = \prod_{i=1}^{N} p(x_i | x_{\pi_i})$$

# DGM for Bayesian Linear Regression




- Model uncertainty over $\mathbf{w}$
- The coefficient vector $\mathbf{w}$ is now a random variable with a prior $p(\mathbf{w}|v) = N(\mathbf{0}, v\mathbf{I})$

# Extra: Why does Naïve Bayes work so well?




- The conditional independence assumptions are strong in NB.
- Why does it work so well in practice?

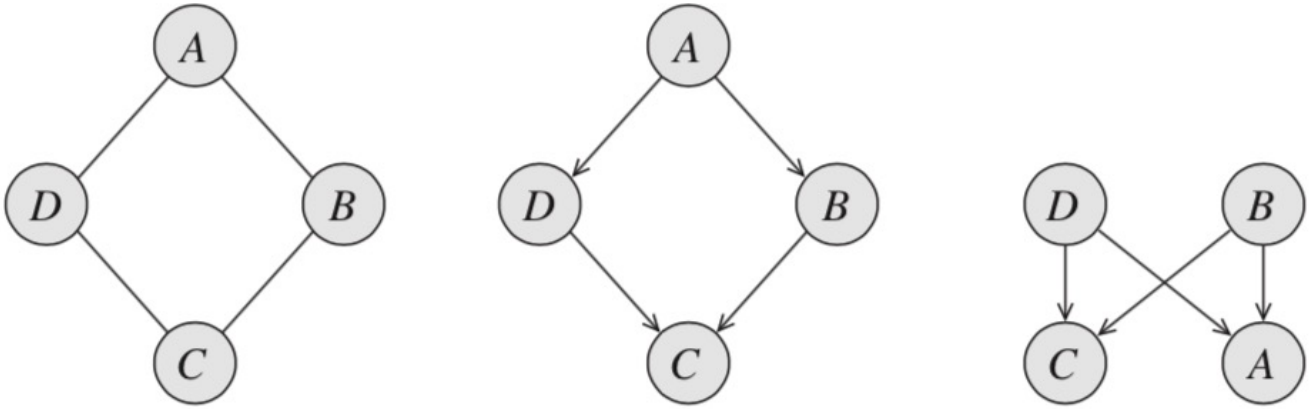
| | MNB | TWCNB | SVM |
|---|---|---|---|
| Industry Sector | 0.582 | 0.923 | 0.934 |
| 20 Newsgroups | 0.848 | 0.861 | 0.862 |
| Reuters (micro) | 0.739 | 0.844 | 0.887 |
| Reuters (macro) | 0.270 | 0.647 | 0.694 |

"Tackling the poor assumptions of Naive Bayes classifiers", *Rennie, J.; Shih, L.; Teevan, J.; Karger,* ICML 2003.

# Questions we want answers to

- Is the Bayesian Network correct/sound?
  - Does a conditional independence identified by d-separation always exist in the distribution? **Yes.** ☺
- Is the Bayesian Network complete?
  - If a conditional independence exists in the distribution, can it always be detected by d-separation? **Almost Yes.** 😐
- How expressive are Bayesian Networks as a modeling language?
  - Can they exactly represent all conditional independencies for a given distribution? **No.** ☹

# Counter-example

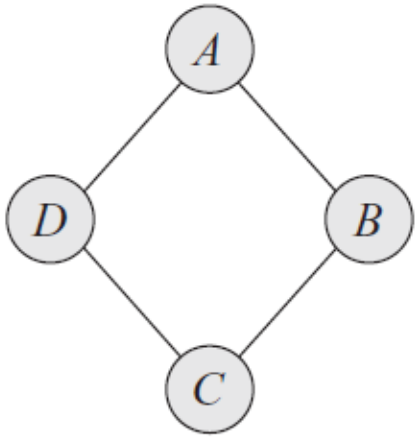



- Can we find a Bayes Net that represents all the conditional independencies in a given probability distribution $P$?
- No. Bayes Nets cannot model certain sets of conditional independence assertions.
- Next week, we will learn about Markov random fields (MRFs) which use undirected links.

# Comparative Semantics

- An example of some CI relationships that can be perfectly modelled by a UGM but not a DGM:

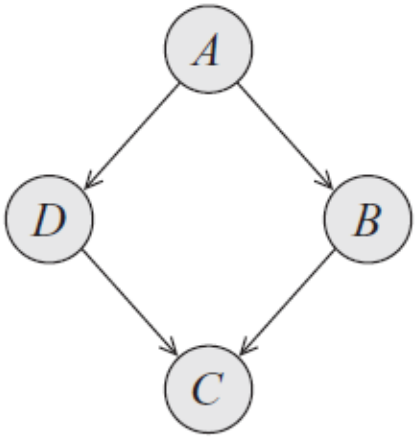
UGM




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

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

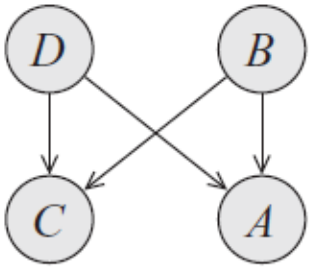
Attempt 1




✔ $A \perp C \mid \{B, D\}$

✖ $B \perp D \mid \{A, C\}$

Attempt 2




✔ $A \perp C \mid \{B, D\}$

✖ $B \perp D \mid \{A, C\}$

Image source: Koller and Friedman 2009

# Course Schedule

| Week | Date | Lecture Topic | Tutorial Topic |
|---|---|---|---|
| 1 | 12 Jan | Introduction to Uncertainty Modeling + Probability Basics | Introduction |
| 2 | 19 Jan | Simple Probabilistic Models | Probability Basics |
| 3 | 26 Jan | Bayesian networks (Directed graphical models) | More Basic Probability |
| 4 | 2 Feb | Markov random Fields (Undirected graphical models) | DGM modelling and d-separation |
| 5 | 9 Feb | Variable elimination and belief propagation | MRF + Sum/Max Product |
| 6 | 16 Feb | Factor graph and the junction tree algorithm | **Quiz 1** |
| - | - | RECESS WEEK | |
| 7 | 2 Mar | Mixture Models and Expectation Maximization (EM) | Linear Gaussian Models |
| 8 | 9 Mar | Hidden Markov Models (HMM) | Probabilistic PCA |
| 9 | **16** Mar | Monte-Carlo Inference (Sampling) | Linear Gaussian Dynamical System |
| 10 | 23 Mar | Variational Inference | MCMC + Sequential VAE |
| 11 | 30 Mar | Inference and Decision-Making (Special Topic) | **Quiz 2** |
| 12 | 6 Apr | Gaussian Processes (Special Topic) | Wellness Day |
| 13 | 13 Apr | **Project Presentations** | Closing |

# Learning Outcomes

- Students should be able to:

1. Explain the concepts of Markov properties (global, local and pairwise) and use it to find all conditional independences in an UGM.

2. Use clique potential functions to parameterize a Markov Random Field, i.e. to represent the joint distribution with clique potential functions.

3. Learn the parameters in a MRF (stochastic) gradient-based methods.

# Acknowledgements


- A lot of slides and content of this lecture are adopted from:

1. "Probabilistic graphical models", Koller and Friedman (Chapter 4)
2. "Machine learning - a probabilistic approach", Kevin Murphy (Chapter 19)
3. “An introduction to probabilistic graphical models”, Michael I. Jordan, 2002 (Section 2.2) http://people.eecs.berkeley.edu/~jordan/prelims/chapter2.pdf
4. "Pattern recognition and machine learning", Christopher Bishop (Chapter 8, Section 8.3).
5. http://www.cs.cmu.edu/~epxing/Class/10708/lectures/lecture3-MRFrepresentation.pdf, Eric Xing
6. Slides from Dr. Lee Gim Hee

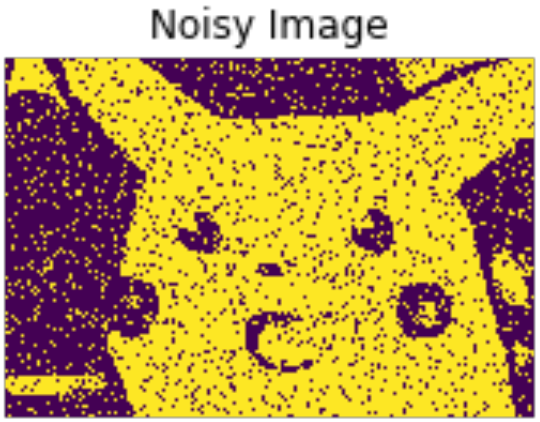
NUS National University of Singapore | School of Computing

# Undirected Graphical Models (Markov Random Fields)

*Examples and Intuition*

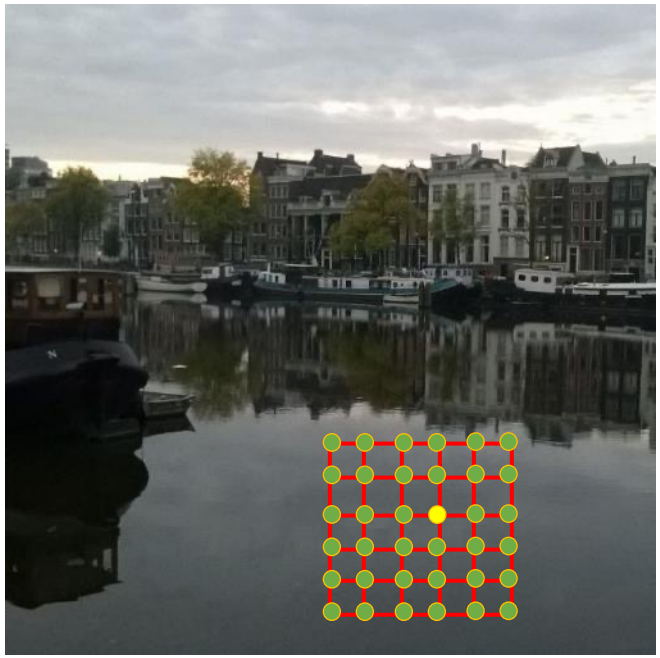
# When directed graphs are strange…

**Examples:**


# When directed graphs are strange...

**Examples:**



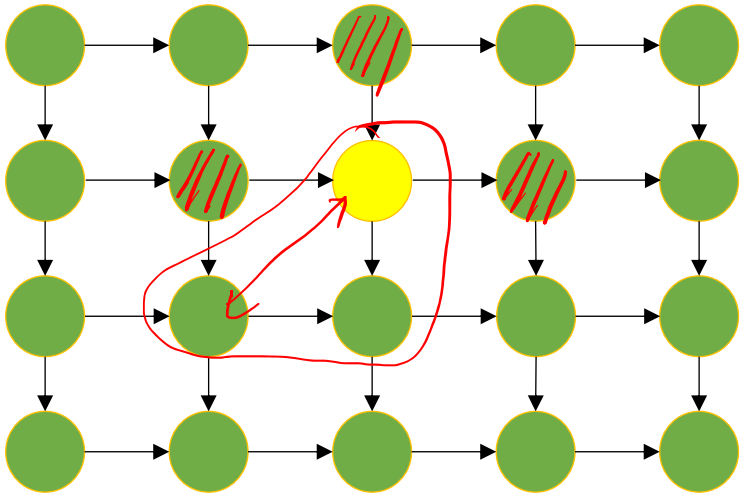
: pixel is labeled as “water”

: is this pixel more likely to be “water” or “sky”?

Photo Source:
G.H. Lee “Amsterdam”

# Recall From Lecture 3

**Examples:**



: pixel is labeled as “water”

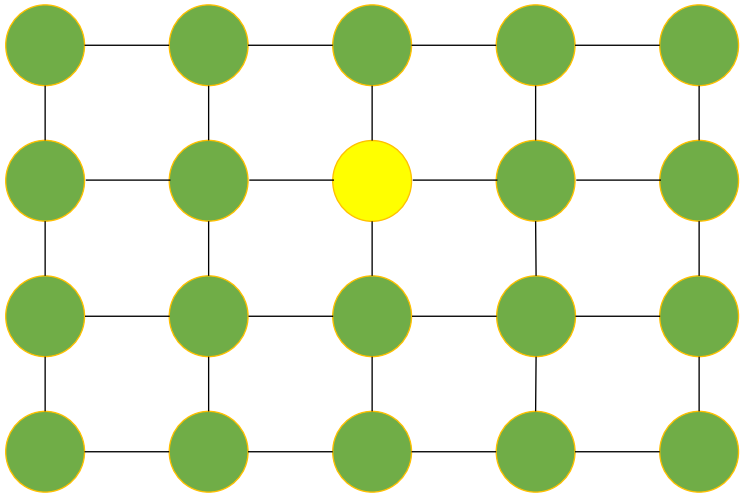
: is this pixel more likely to be “water” or “sky”?

**What are the conditional independence assertions?**

**This is a little weird!**

# Recall From Lecture 3

**Examples:**



: pixel is labeled as “water”

: is this pixel more likely to be “water” or “sky”?

**What if we gave up on the arrows?**

# Undirected Graphical Models:
## In a nutshell

- An alternative to DGMs is to use an Undirected Graphical model (UGM), also called a Markov Random Field (MRF) or Markov network.

- Formally, an UGM is a graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$, where:
  - $\mathcal{V}$ is a set of nodes that are in one-to-one correspondence with a set of random variables.
  - $\mathcal{E}$ is a set of undirected edges.

# Undirected Graphical Models:
## In a nutshell

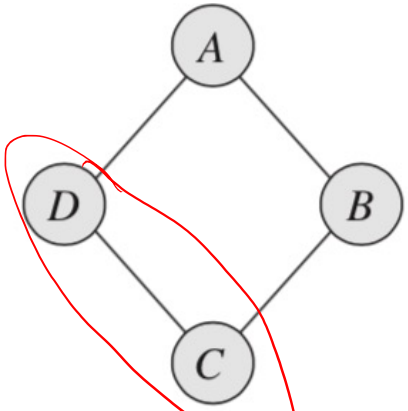
- Parameterization is achieved via factors.
- A factor $\varphi(\mathcal{C})$ is a function that maps a set of random variables $\mathcal{C} = \{X, \dots, Z\}$ to a real number.
  - Restrict: non-negative factors only ($\varphi(\mathcal{C})$ only maps to non-negative numbers)
- The factorization is:

$$p(x_1, \dots, x_N) = \frac{1}{Z} \prod_{j=1}^{M} \varphi_j(\mathcal{C}_j)$$

Compare to DGM:

$$p(x_1, \dots, x_N) = \prod_{i=1}^{N} p(x_i | x_{\pi_i})$$

# Factor Example




$\phi_1(A, B)$

| | | |
|---|---|---|
| $a^0$ | $b^0$ | 30 |
| $a^0$ | $b^1$ | 5 |
| $a^1$ | $b^0$ | 1 |
| $a^1$ | $b^1$ | 10 |

$\phi_2(B, C)$

| | | |
|---|---|---|
| $b^0$ | $c^0$ | 100 |
| $b^0$ | $c^1$ | 1 |
| $b^1$ | $c^0$ | 1 |
| $b^1$ | $c^1$ | 100 |

$\phi_3(C, D)$

| | | |
|---|---|---|
| $c^0$ | $d^0$ | 1 |
| $c^0$ | $d^1$ | 100 |
| $c^1$ | $d^0$ | 100 |
| $c^1$ | $d^1$ | 1 |

$\phi_4(D, A)$

| | | |
|---|---|---|
| $d^0$ | $a^0$ | 100 |
| $d^0$ | $a^1$ | 1 |
| $d^1$ | $a^0$ | 1 |
| $d^1$ | $a^1$ | 100 |

Image source: Koller and Friedman 2009

# UGMs Vs DGMs

Advantages of UGMs over DGMs are:

1. No edge orientations, hence more natural for some problems such as image analysis and spatial statistics.

2. Discriminative UGMs (aka conditional random fields, or CRFs) work better than discriminative DGMs (more on discriminative models later).

Disadvantages of UGMs over DGMs are:

1. The parameters are less interpretable and less modular (details later).
2. Parameter estimation is computationally more expensive.

# UGM Misconceptions

- Some misconceptions about UGMs / MRFs:
  - ✗ Factors always represent marginal/conditional distributions.
  - ✗ Undirected graphs are "richer" (can represent more conditional independencies) compared to DGMs
  - ✗ UGMs specify a unique factorization.
- These are incorrect!
- We will learn the right way to interpret UGMs.
- Just like DGMs, UGMs encode a set of conditional independence assertions.

NUS National University of Singapore | School of Computing

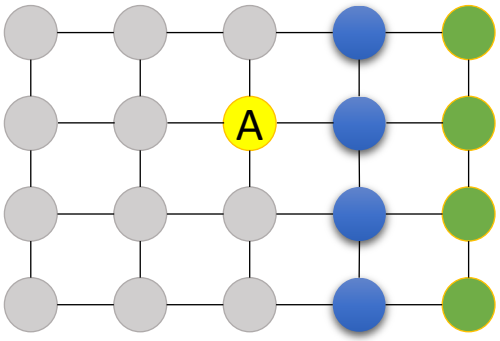
# Conditional Independence in MRFs

*Global, Local, and Pairwise Markov Properties*

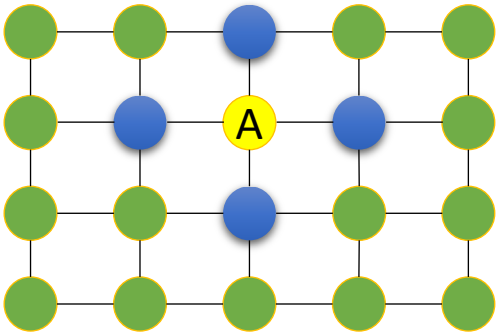
# Conditional Independence

- 3 Markov Properties of UGMs:
  1. Global Markov Property
  2. Local Markov Property
  3. Pairwise Markov Property

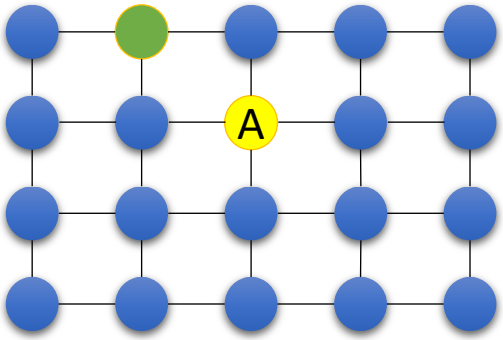
**Global**




**Local**




**Pairwise**

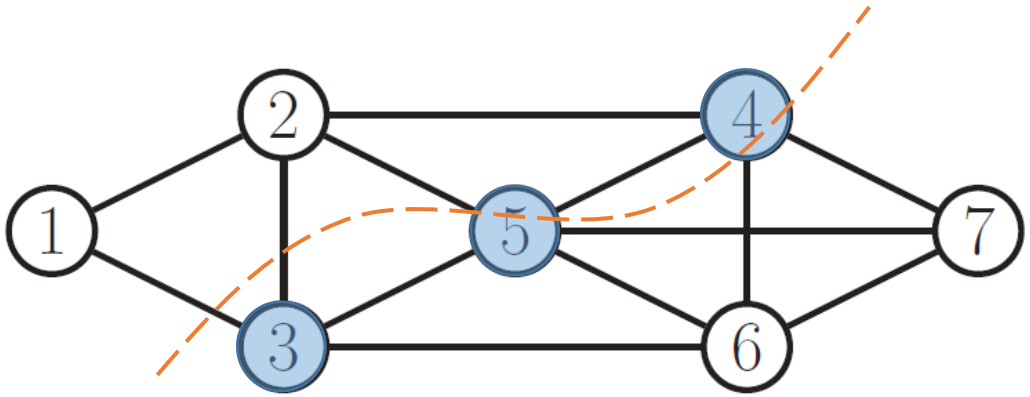


# Conditional Independence

## 1. Global Markov Property

- Given the sets of nodes $A, B$ and $C$, $X_A \perp X_B \mid X_C$ if and only if $C$ separates $A$ from $B$ in the graph $\mathcal{G}$.
- This means that there are no trails connecting any node in $A$ to any node in $B$ when we remove all nodes in $C$.

**Example:**




$$\{X_1, X_2\} \perp \{X_6, X_7\} \mid \{X_3, X_4, X_5\}$$

Image Source: Modified from "Machine Learning – A Probabilistic Perspective", Kevin Murphy

# Conditional Independence

## 1. Global Markov Property

- Given the sets of nodes $A, B$ and $C$, $X_A \perp X_B \mid X_C$ if and only if $C$ separates $A$ from $B$ in the graph $\mathcal{G}$.
- This means that there are no trails connecting any node in $A$ to any node in $B$ when we remove all nodes in $C$.

**Example:**

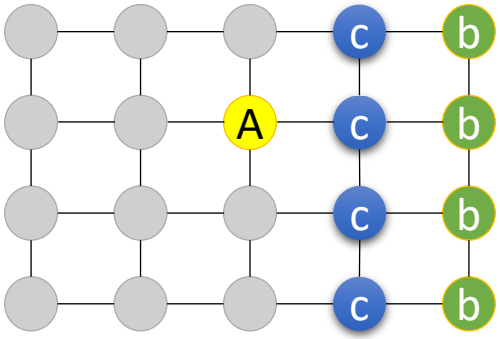
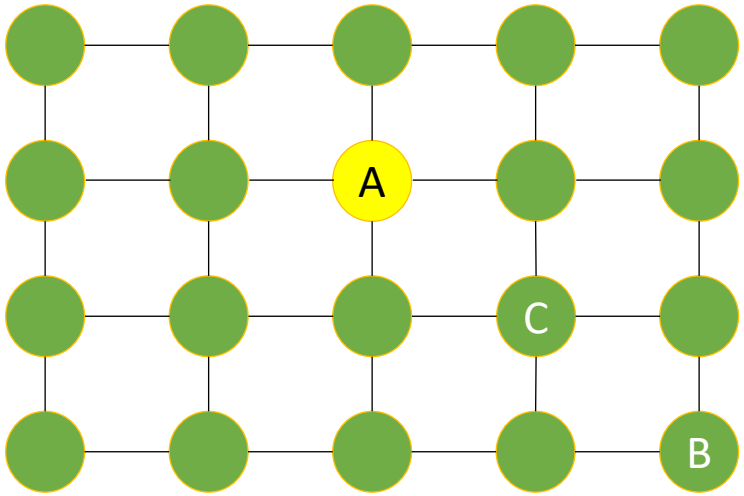



Image Source: Modified from "Machine Learning – A Probabilistic Perspective", Kevin Murphy

# Global Markov Property

**Example:**

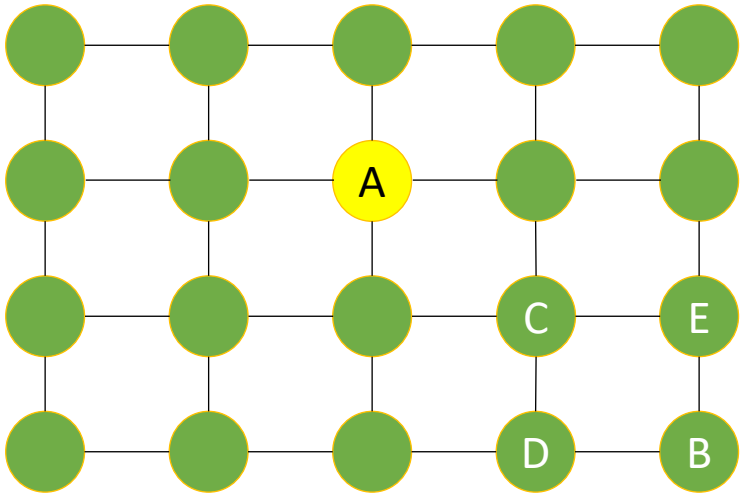



Given the sets of nodes $A, B$ and $C$, $X_A \perp X_B \mid X_C$ if and only if $C$ separates $A$ from $B$ in the graph $\mathcal{G}$.

**Is $A \perp B \mid \{C\}$?** **No**

# Global Markov Property

**Example:**




Given the sets of nodes $A, B$ and $C$, $X_A \perp X_B \mid X_C$ if and only if $C$ separates $A$ from $B$ in the graph $\mathcal{G}$.

**Is $A \perp B \mid \{C, D, E\}$? Yes**

# Conditional Independence

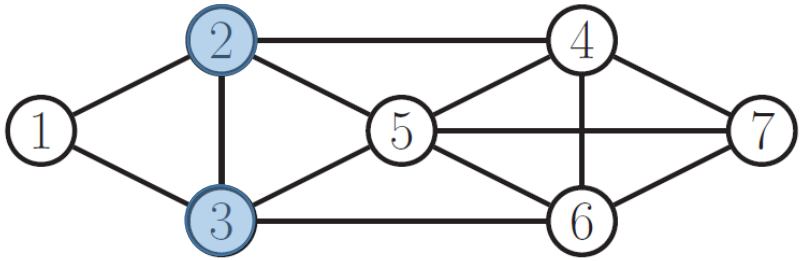
## 2. Local Markov Property

- The Markov blanket of $X_s$ denoted $\text{mb}(X_s)$ is the set of nodes that renders a node $X_s$ conditionally independent of all the other nodes in $\mathcal{G}$ : All other nodes in G

$$X_s \perp \mathcal{V} \setminus \{\text{mb}(X_s), X_s\} \mid \text{mb}(X_s)$$

- The Markov Blanket in a UGM is the set of its immediate neighbours.

**Example:**

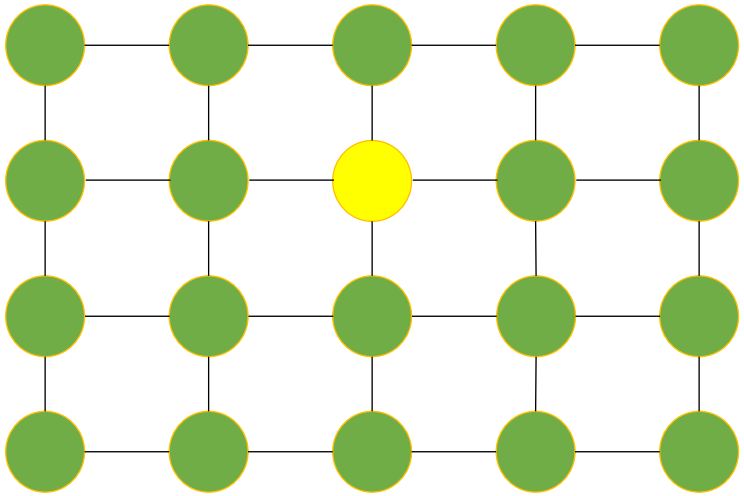



$$\text{mb}(X_1) = \{X_2, X_3\}, \text{ i.e. } X_1 \perp \{X_4, X_5, X_6, X_7\} \mid \{X_2, X_3\}$$

Image Source: Modified from "Machine Learning – A Probabilistic Perspective", Kevin Murphy

# Markov Blanket in UGM

**Example:**



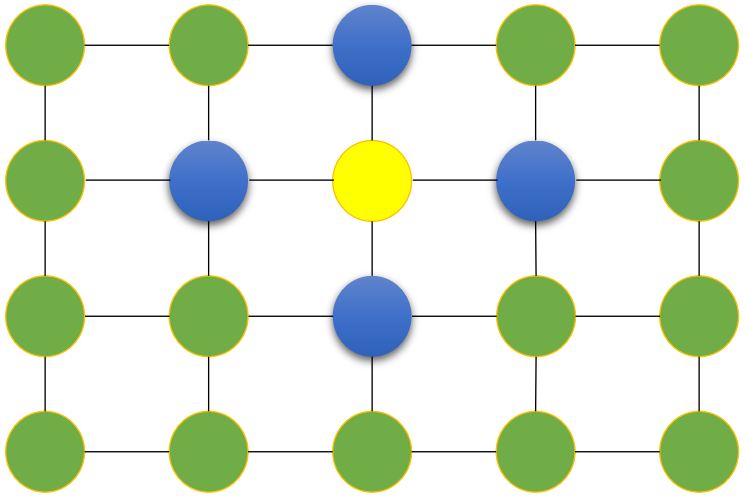
The Markov blanket of $X_s$ denoted $\text{mb}(X_s)$ is the set of nodes that renders a node $X_s$ conditionally independent of all the other nodes in $\mathcal{G}$ :

$$X_s \perp \mathcal{V} \setminus \{\text{mb}(X_s), X_s\} \mid \text{mb}(X_s)$$

**What is the Markov Blanket of ● ?**

# Markov Blanket in UGM

**Example:**



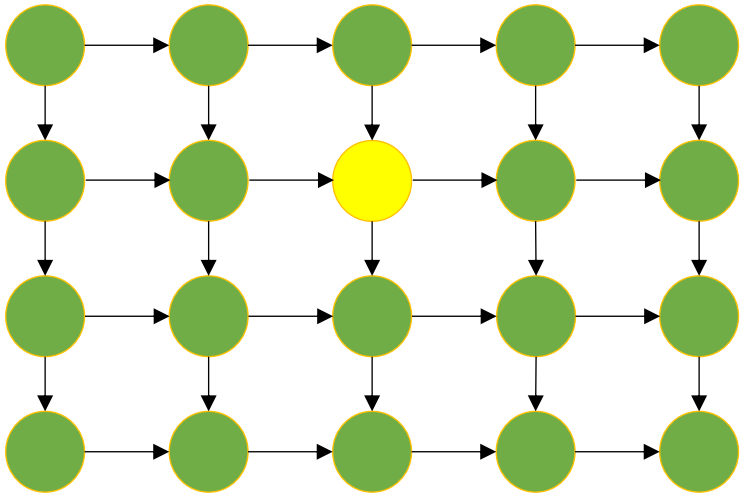
The Markov blanket of $X_s$ denoted $\text{mb}(X_s)$ is the set of nodes that renders a node $X_s$ conditionally independent of all the other nodes in $\mathcal{G}$ :

$$X_s \perp \mathcal{V} \setminus \{\text{mb}(X_s), X_s\} \mid \text{mb}(X_s)$$

The Markov Blanket are the **blue nodes** due to the Global Markov Property

# Markov Blanket in a DGM?

**Examples:**



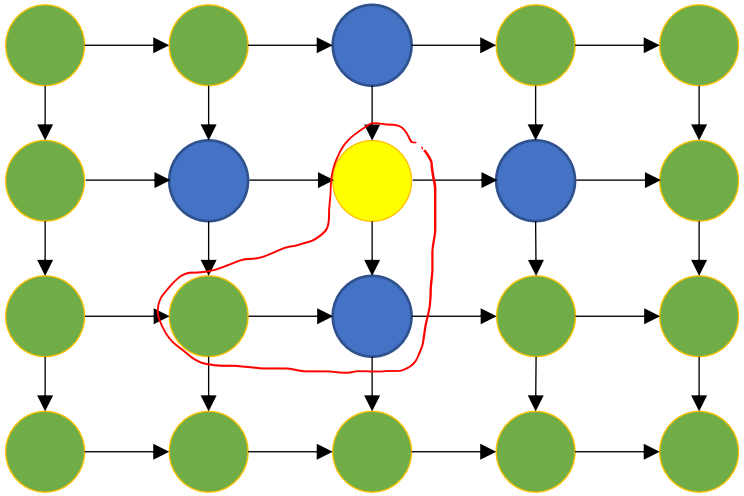
The Markov blanket of $X_s$ denoted $\text{mb}(X_s)$ is the set of nodes that renders a node $X_s$ conditionally independent of all the other nodes in $\mathcal{G}$ :

$$X_s \perp \mathcal{V} \setminus \{\text{mb}(X_s), X_s\} \mid \text{mb}(X_s)$$

**What is the Markov Blanket of ● ?**

# Markov Blanket in a DGM?

**Examples:**



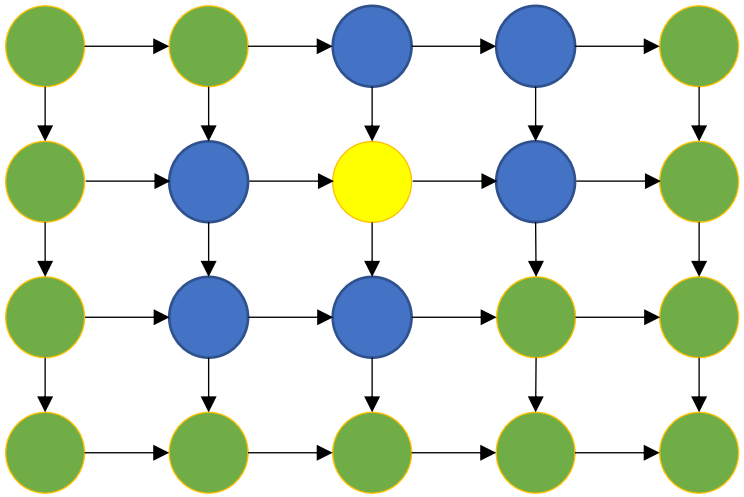
The Markov blanket of $X_s$ denoted $\text{mb}(X_s)$ is the set of nodes that renders a node $X_s$ conditionally independent of all the other nodes in $\mathcal{G}$ :

$$X_s \perp \mathcal{V} \setminus \{\text{mb}(X_s), X_s\} \mid \text{mb}(X_s)$$

Is this correct?

# Markov Blanket in a DGM?

**Examples:**



The Markov blanket of $X_s$ denoted $\mathrm{mb}(X_s)$ is the set of nodes that renders a node $X_s$ conditionally independent of all the other nodes in $\mathcal{G}$ :

$$X_s \perp \mathcal{V} \setminus \{\mathrm{mb}(X_s), X_s\} \mid \mathrm{mb}(X_s)$$

Markov Blankets for a node $X_s$ in a DGM is the set of the node's parents, children and co-parents (other parents of children)

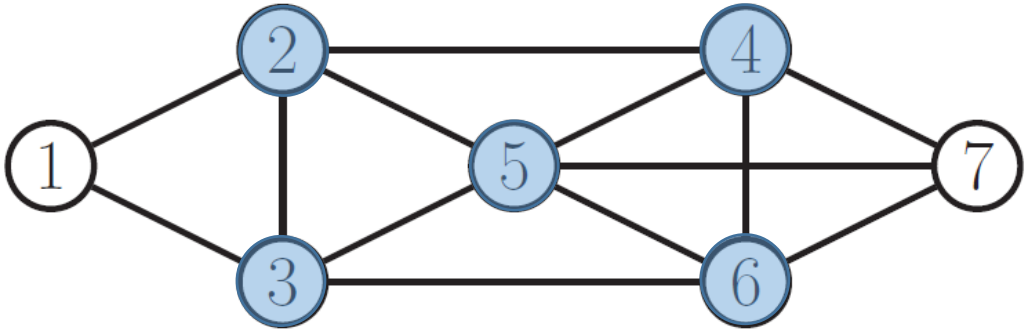
# Conditional Independence

## 3. Pairwise Markov Property

- Two nodes $X_s$ and $X_t$ are conditionally independent given the rest if there is no direct edge between them:

$$X_s \perp X_t \mid \mathcal{V} \setminus \{X_s, X_t\}, \text{ where } \mathcal{E}_{st} = \emptyset$$

**Example:**

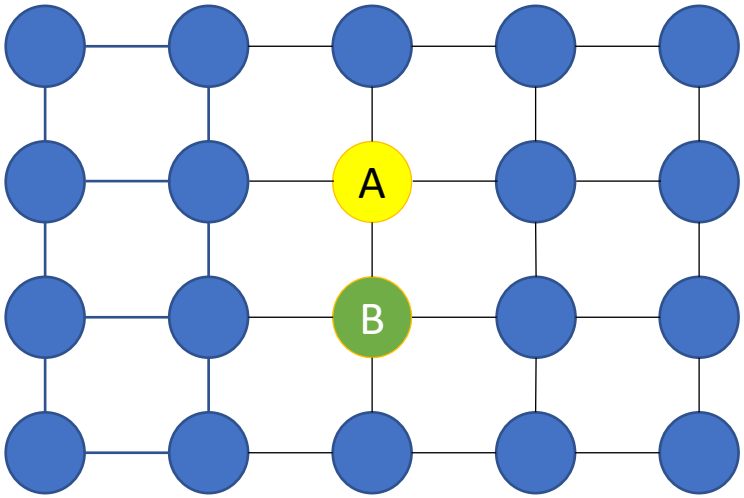



$$X_1 \perp X_7 \mid \{X_2, X_3, X_4, X_5, X_6\}$$

Image Source: Modified from "Machine Learning – A Probabilistic Perspective", Kevin Murphy

# Pairwise Markov Property

**Example:**



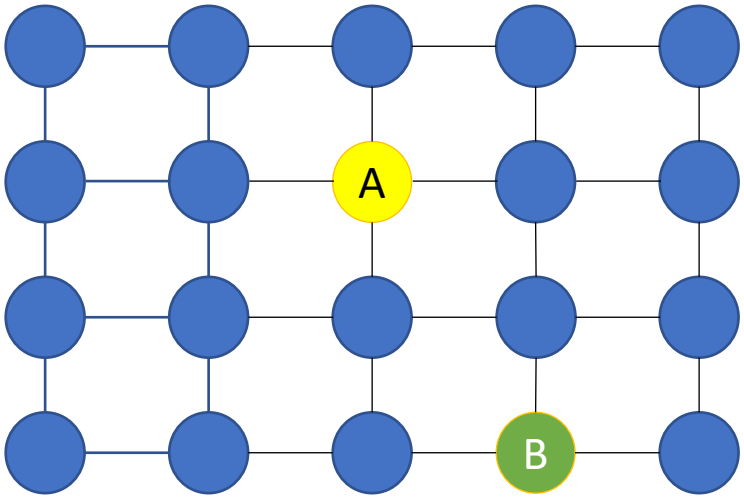

Two nodes $X_s$ and $X_t$ are conditionally independent given the rest if there is no direct edge between them:

$$X_s \perp X_t \mid \mathcal{V} \setminus \{X_s, X_t\}, \text{ where } \mathcal{E}_{st} = \emptyset$$

**Is $A \perp B \mid V \setminus \{A, B\}$? No**

# Pairwise Markov Property

**Example:**



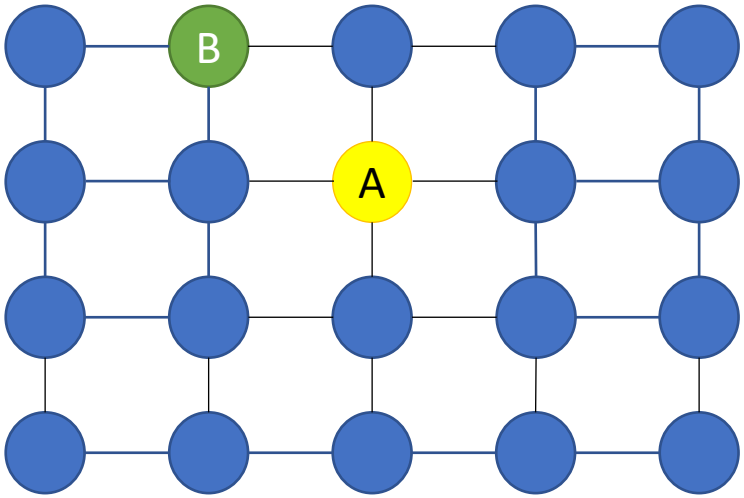

Two nodes $X_s$ and $X_t$ are conditionally independent given the rest if there is no direct edge between them:

$$X_s \perp X_t \mid \mathcal{V} \setminus \{X_s, X_t\}, \text{ where } \mathcal{E}_{st} = \emptyset$$

**Is $A \perp B \mid V \setminus \{A, B\}$? Yes**

# Pairwise Markov Property

**Example:**




Two nodes $X_s$ and $X_t$ are conditionally independent given the rest if there is no direct edge between them:

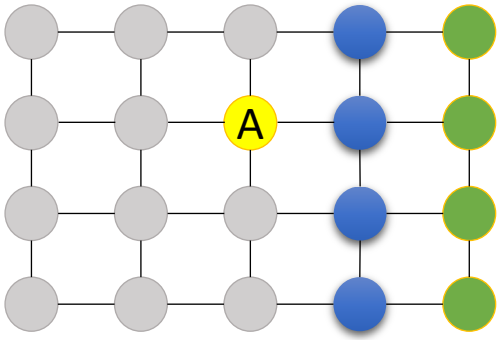
$$X_s \perp X_t \mid \mathcal{V} \setminus \{X_s, X_t\}, \text{ where } \mathcal{E}_{st} = \emptyset$$

**Is $A \perp B \mid V \setminus \{A, B\}$? Yes**

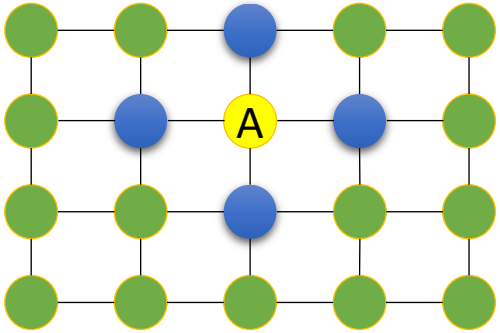
# Conditional Independence

- 3 Markov Properties of UGMs:
  1. Global Markov Property
  2. Local Markov Property
  3. Pairwise Markov Property

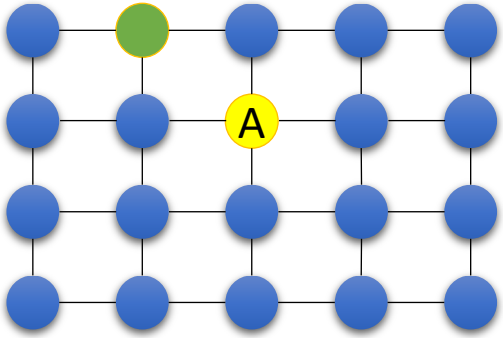
**Global**




**Local**



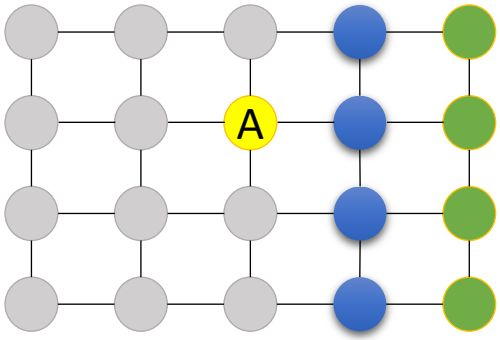

**Pairwise**



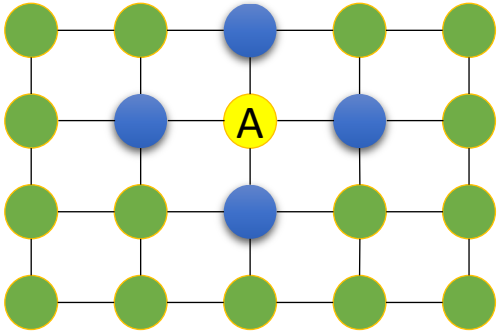
# Conditional Independence

- 3 Markov Properties of UGMs:
  1. Global Markov Property
  2. Local Markov Property
  3. Pairwise Markov Property
- The three properties are interrelated.

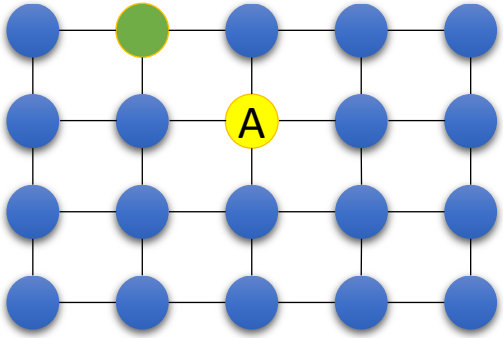
**Global**




**Local**




**Pairwise**



# Global-Local-Pairwise Markov

$$G \Longrightarrow L \Longrightarrow P$$

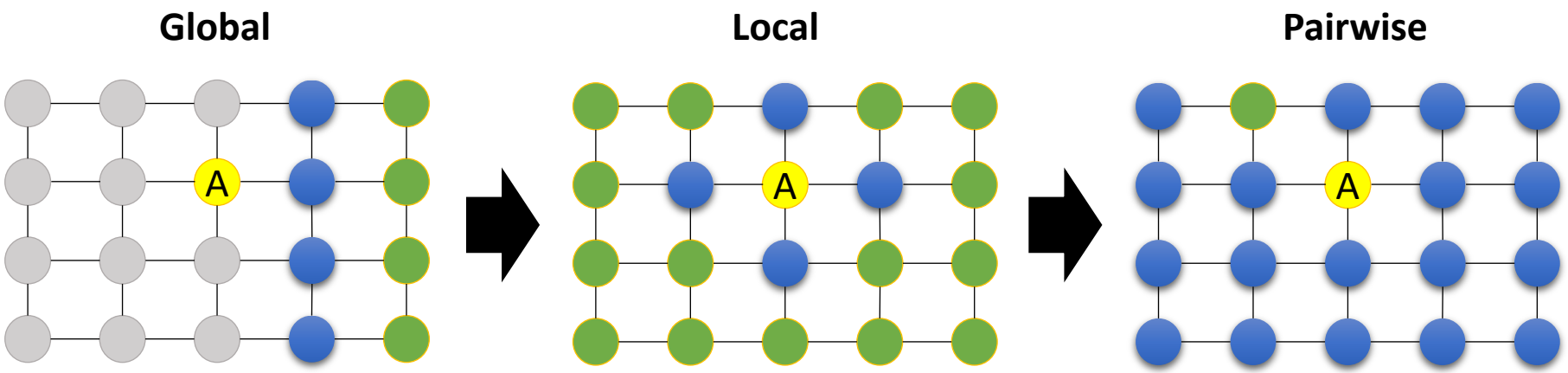
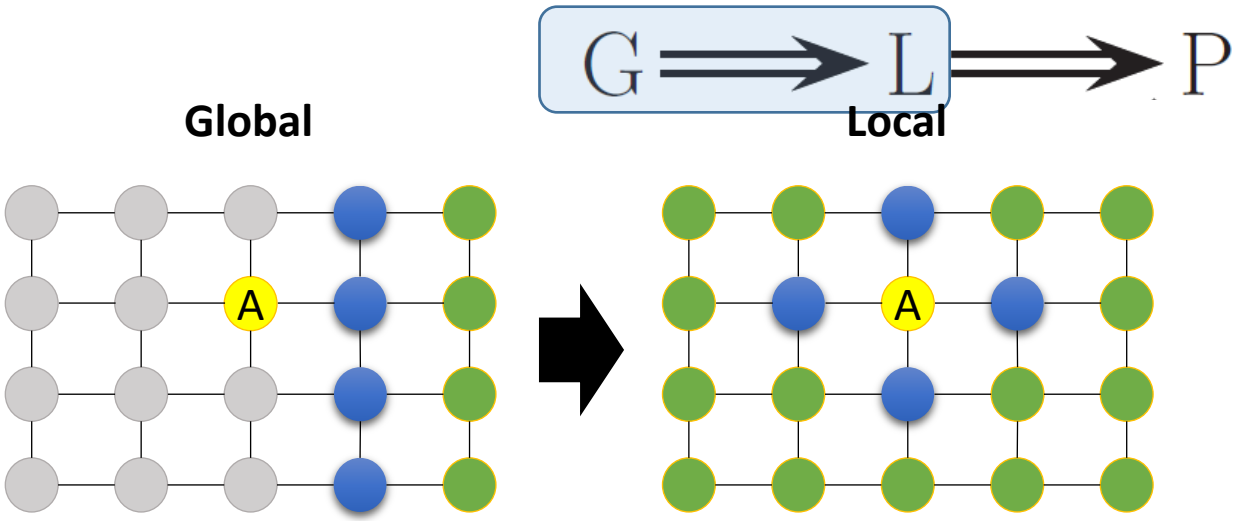



Image Source: “Machine Learning – A Probabilistic Perspective”, Kevin Murphy

# Global Markov implies Local Markov

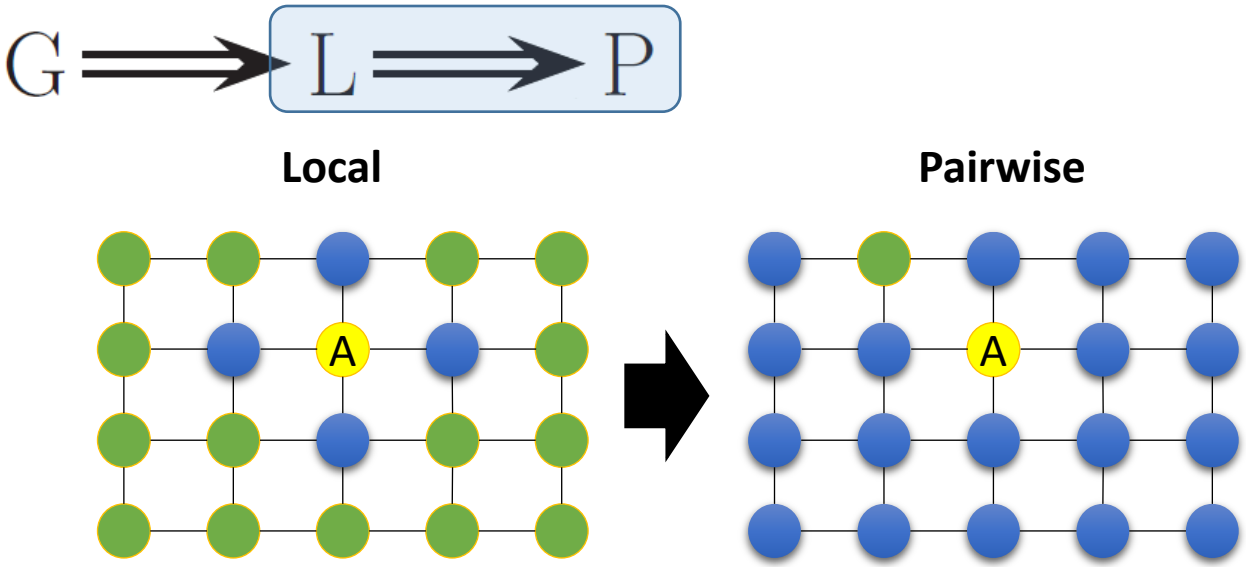



The global Markov property implies the local Markov property: this is the case when the sets $X_A = X_s,\ X_C = \mathrm{mb}(X_s)$, and $X_B = \{\mathcal{V} \setminus \{\mathrm{mb}(X_s), X_s\}$.

$$X_A \perp X_B \mid X_C \ \Rightarrow\ X_s \perp \mathcal{V} \setminus \{\mathrm{mb}(X_s), X_s\} \mid \mathrm{mb}(X_s)$$

Image Source: "Machine Learning – A Probabilistic Perspective", Kevin Murphy

# Local Markov implies Pairwise Markov

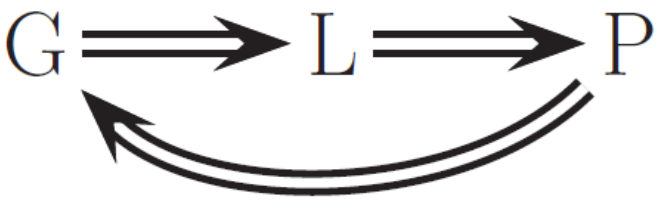



Given any node $X_t$ that is not adjacent to the node $X_s$, it follows from local Markov property that:

$$X_s \perp X_{\mathcal{V} \setminus \{\mathrm{mb}(X_s), X_s\}} \mid \mathrm{mb}(X_s)$$

This implies $X_s \perp X_t \mid X_{\mathcal{V} \setminus \{X_s, X_t\}}$, i.e. pairwise Markov property.

# Conditional Independence




$$p(x) > 0$$

- Easy to check that global Markov implies local Markov which implies pairwise Markov.
- What is less obvious, but true (assuming positive distributions $p(\mathbf{x}) > 0$ for all $\mathbf{x}$), is that pairwise Markov implies global Markov.
  - Proof by induction in Koller and Friedman (**Thm 4.4**)

Image Source: "Machine Learning – A Probabilistic Perspective", Kevin Murphy

# Conditional Independence

- 3 Markov Properties of UGMs:
  1. Global Markov Property
  2. Local Markov Property
  3. Pairwise Markov Property
- The three properties are interrelated.
- For positive distributions $p(\mathbf{x}) > 0$, the three Markov properties are equivalent.

NUS National University of Singapore | School of Computing

# Parameterization of MRFs

*Factors and Gibbs distributions*

# Parameterization of MRFs

- As in the case of DGMs, we would like to obtain a local parameterization for UGMs.
- We have seen earlier that for *DGMs*:
  - Parameterization was based on local conditional probabilities of a node and its parents, i.e. $p(x_i|x_{\pi_i})$.
  - Joint probability is a product of local conditional probabilities as a result of the chain rule, i.e.

$$p(x_1, \ldots, x_N) = \prod_{i=1}^{N} p(x_i|x_{\pi_i})$$

# Parameterization of MRFs

- Difficult to do local parameterization based on conditional probabilities since no topological ordering associated with UGMs.

- It turns out that its better to abandon conditional probabilities altogether, and use some functions instead.

# Undirected Graphical Models:
## In a nutshell

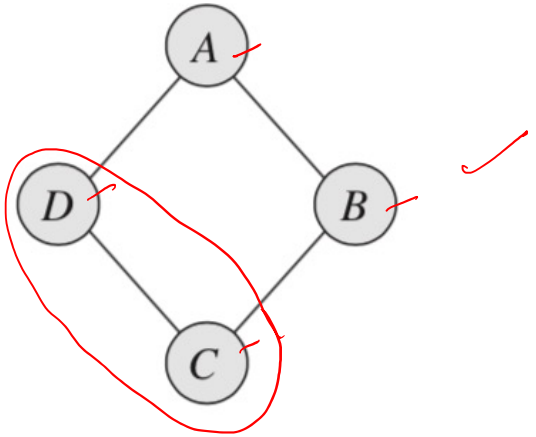
- Parameterization is achieved via factors.
- A factor $\varphi(\mathcal{C})$ is a function that maps a set of random variables $\mathcal{C} = \{X, \ldots, Z\}$ to a real number.
  - Restrict: non-negative factors only ($\varphi(\mathcal{C})$ only maps to non-negative numbers)
- The factorization is:

$$p(x_1, \ldots, x_N) = \frac{1}{Z} \prod_{j=1}^{M} \varphi_j(\mathcal{C}_j)$$

Compare to DGM:

$$p(x_1, \ldots, x_N) = \prod_{i=1}^{N} p(x_i | x_{\pi_i})$$

# A Specific Example




$\varphi(A,B)$

| | | |
|---|---|---|
| $a^0$ | $b^0$ | 30 |
| $a^0$ | $b^1$ | 5 |
| $a^1$ | $b^0$ | 1 |
| $a^1$ | $b^1$ | 10 |

$\varphi(B,C)$

| | | |
|---|---|---|
| $b^0$ | $c^0$ | 100 |
| $b^0$ | $c^1$ | 1 |
| $b^1$ | $c^0$ | 1 |
| $b^1$ | $c^1$ | 100 |

$\varphi(C,D)$

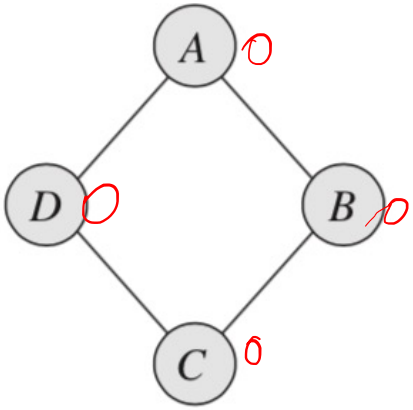
| | | |
|---|---|---|
| $c^0$ | $d^0$ | 1 |
| $c^0$ | $d^1$ | 100 |
| $c^1$ | $d^0$ | 100 |
| $c^1$ | $d^1$ | 1 |

$\varphi(D,A)$

| | | |
|---|---|---|
| $d^0$ | $a^0$ | 100 |
| $d^0$ | $a^1$ | 1 |
| $d^1$ | $a^0$ | 1 |
| $d^1$ | $a^1$ | 100 |

Image source: Koller and Friedman 2009

# A Specific Example




| Assignment | | | | Unnormalized |
|---|---|---|---|---|
| $a^0$ | $b^0$ | $c^0$ | $d^0$ | 300,000 |
| $a^0$ | $b^0$ | $c^0$ | $d^1$ | 300,000 |
| $a^0$ | $b^0$ | $c^1$ | $d^0$ | 300,000 |
| $a^0$ | $b^0$ | $c^1$ | $d^1$ | 30 |

$\varphi(A,B)$

| | | |
|---|---|---|
| $a^0$ | $b^0$ | 30 |
| $a^0$ | $b^1$ | 5 |
| $a^1$ | $b^0$ | 1 |
| $a^1$ | $b^1$ | 10 |

$\varphi(B,C)$

| | | |
|---|---|---|
| $b^0$ | $c^0$ | 100 |
| $b^0$ | $c^1$ | 1 |
| $b^1$ | $c^0$ | 1 |
| $b^1$ | $c^1$ | 100 |

$\varphi(C,D)$

| | | |
|---|---|---|
| $c^0$ | $d^0$ | 1 |
| $c^0$ | $d^1$ | 100 |
| $c^1$ | $d^0$ | 100 |
| $c^1$ | $d^1$ | 1 |

$\varphi(D,A)$

| | | |
|---|---|---|
| $d^0$ | $a^0$ | 100 |
| $d^0$ | $a^1$ | 1 |
| $d^1$ | $a^0$ | 1 |
| $d^1$ | $a^1$ | 100 |

Image source: Koller and Friedman 2009

| $Assignment$ | | | | $Unnormalized$ |
|---|---|---|---|---|
| $a^0$ | $b^0$ | $c^0$ | $d^0$ | $300,000$ |
| $a^0$ | $b^0$ | $c^0$ | $d^1$ | $300,000$ |
| $a^0$ | $b^0$ | $c^1$ | $d^0$ | $300,000$ |
| $a^0$ | $b^0$ | $c^1$ | $d^1$ | $30$ |
| $a^0$ | $b^1$ | $c^0$ | $d^0$ | $500$ |
| $a^0$ | $b^1$ | $c^0$ | $d^1$ | $500$ |
| $a^0$ | $b^1$ | $c^1$ | $d^0$ | $5,000,000$ |
| $a^0$ | $b^1$ | $c^1$ | $d^1$ | $500$ |
| $a^1$ | $b^0$ | $c^0$ | $d^0$ | $100$ |
| $a^1$ | $b^0$ | $c^0$ | $d^1$ | $1,000,000$ |
| $a^1$ | $b^0$ | $c^1$ | $d^0$ | $100$ |
| $a^1$ | $b^0$ | $c^1$ | $d^1$ | $100$ |
| $a^1$ | $b^1$ | $c^0$ | $d^0$ | $10$ |
| $a^1$ | $b^1$ | $c^0$ | $d^1$ | $100,000$ |
| $a^1$ | $b^1$ | $c^1$ | $d^0$ | $100,000$ |
| $a^1$ | $b^1$ | $c^1$ | $d^1$ | $100,000$ |

Have to normalize!

Z is our normalizing constant

In this case:

$Z = 7{,}201{,}840$

Image source: Koller and Friedman 2009

| Assignment | | | | Unnormalized | Normalized |
|---|---|---|---|---|---|
| $a^0$ | $b^0$ | $c^0$ | $d^0$ | $300,000$ | $0.04$ |
| $a^0$ | $b^0$ | $c^0$ | $d^1$ | $300,000$ | $0.04$ |
| $a^0$ | $b^0$ | $c^1$ | $d^0$ | $300,000$ | $0.04$ |
| $a^0$ | $b^0$ | $c^1$ | $d^1$ | $30$ | $4.1 \cdot 10^{-6}$ |
| $a^0$ | $b^1$ | $c^0$ | $d^0$ | $500$ | $6.9 \cdot 10^{-5}$ |
| $a^0$ | $b^1$ | $c^0$ | $d^1$ | $500$ | $6.9 \cdot 10^{-5}$ |
| $a^0$ | $b^1$ | $c^1$ | $d^0$ | $5,000,000$ | $0.69$ |
| $a^0$ | $b^1$ | $c^1$ | $d^1$ | $500$ | $6.9 \cdot 10^{-5}$ |
| $a^1$ | $b^0$ | $c^0$ | $d^0$ | $100$ | $1.4 \cdot 10^{-5}$ |
| $a^1$ | $b^0$ | $c^0$ | $d^1$ | $1,000,000$ | $0.14$ |
| $a^1$ | $b^0$ | $c^1$ | $d^0$ | $100$ | $1.4 \cdot 10^{-5}$ |
| $a^1$ | $b^0$ | $c^1$ | $d^1$ | $100$ | $1.4 \cdot 10^{-5}$ |
| $a^1$ | $b^1$ | $c^0$ | $d^0$ | $10$ | $1.4 \cdot 10^{-6}$ |
| $a^1$ | $b^1$ | $c^0$ | $d^1$ | $100,000$ | $0.014$ |
| $a^1$ | $b^1$ | $c^1$ | $d^0$ | $100,000$ | $0.014$ |
| $a^1$ | $b^1$ | $c^1$ | $d^1$ | $100,000$ | $0.014$ |

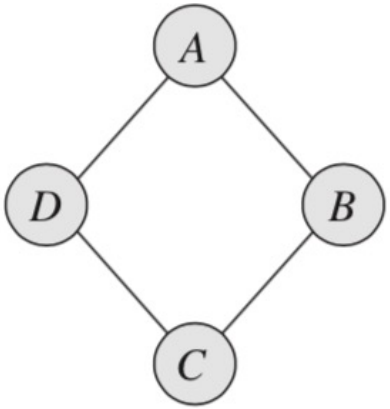


Image source: Koller and Friedman 2009

# Question

Consider pairwise factor $\varphi(A, B)$. Is the factor proportional to:

A. The marginal probability $p(A, B)$
B. The conditional probability $p(A|B)$
C. All of the above
D. None of the above




$\varphi(A, B)$

| | | |
|---|---|---|
| $a^0$ | $b^0$ | 30 |
| $a^0$ | $b^1$ | 5 |
| $a^1$ | $b^0$ | 1 |
| $a^1$ | $b^1$ | 10 |

Image source: Koller and Friedman 2009

# Question

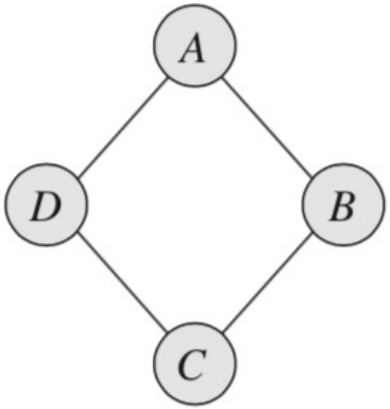
Consider pairwise factor $\varphi(A, B)$. Is the factor proportional to:

A. The marginal probability $p(A, B)$

B. The conditional probability $p(A|B)$

C. All of the above

**D. None of the above**




$\varphi(A, B)$

| | | |
|---|---|---|
| $a^0$ | $b^0$ | 30 |
| $a^0$ | $b^1$ | 5 |
| $a^1$ | $b^0$ | 1 |
| $a^1$ | $b^1$ | 10 |

| $a$ | $b$ | $c$ | $d$ | Unnormalized | Normalized |
|---|---|---|---|---|---|
| $a^0$ | $b^0$ | $c^0$ | $d^0$ | 300,000 | 0.04 |
| $a^0$ | $b^0$ | $c^0$ | $d^1$ | 300,000 | 0.04 |
| $a^0$ | $b^0$ | $c^1$ | $d^0$ | 300,000 | 0.04 |
| $a^0$ | $b^0$ | $c^1$ | $d^1$ | 30 | $4.1 \cdot 10^{-6}$ |
| $a^0$ | $b^1$ | $c^0$ | $d^0$ | 500 | $6.9 \cdot 10^{-5}$ |
| $a^0$ | $b^1$ | $c^0$ | $d^1$ | 500 | $6.9 \cdot 10^{-5}$ |
| $a^0$ | $b^1$ | $c^1$ | $d^0$ | 5,000,000 | 0.69 |
| $a^0$ | $b^1$ | $c^1$ | $d^1$ | 500 | $6.9 \cdot 10^{-5}$ |
| $a^1$ | $b^0$ | $c^0$ | $d^0$ | 100 | $1.4 \cdot 10^{-5}$ |
| $a^1$ | $b^0$ | $c^0$ | $d^1$ | 1,000,000 | 0.14 |
| $a^1$ | $b^0$ | $c^1$ | $d^0$ | 100 | $1.4 \cdot 10^{-5}$ |
| $a^1$ | $b^0$ | $c^1$ | $d^1$ | 100 | $1.4 \cdot 10^{-5}$ |
| $a^1$ | $b^1$ | $c^0$ | $d^0$ | 10 | $1.4 \cdot 10^{-6}$ |
| $a^1$ | $b^1$ | $c^0$ | $d^1$ | 100,000 | 0.014 |
| $a^1$ | $b^1$ | $c^1$ | $d^0$ | 100,000 | 0.014 |
| $a^1$ | $b^1$ | $c^1$ | $d^1$ | 100,000 | 0.014 |

$p(a^0, b^0)$

$= \sum_{c,d} p(a^0, b^0, c, d)$

≈0.13



Image source: Koller and Friedman 2009

| Assignment | | | | Unnormalized | Normalized |
|---|---|---|---|---|---|
| $a^0$ | $b^0$ | $c^0$ | $d^0$ | 300,000 | 0.04 |
| $a^0$ | $b^0$ | $c^0$ | $d^1$ | 300,000 | 0.04 |
| $a^0$ | $b^0$ | $c^1$ | $d^0$ | 300,000 | 0.04 |
| $a^0$ | $b^0$ | $c^1$ | $d^1$ | 30 | $4.1 \cdot 10^{-6}$ |
| $a^0$ | $b^1$ | $c^0$ | $d^0$ | 500 | $6.9 \cdot 10^{-5}$ |
| $a^0$ | $b^1$ | $c^0$ | $d^1$ | 500 | $6.9 \cdot 10^{-5}$ |
| $a^0$ | $b^1$ | $c^1$ | $d^0$ | 5,000,000 | 0.69 |
| $a^0$ | $b^1$ | $c^1$ | $d^1$ | 500 | $6.9 \cdot 10^{-5}$ |
| $a^1$ | $b^0$ | $c^0$ | $d^0$ | 100 | $1.4 \cdot 10^{-5}$ |
| $a^1$ | $b^0$ | $c^0$ | $d^1$ | 1,000,000 | 0.14 |
| $a^1$ | $b^0$ | $c^1$ | $d^0$ | 100 | $1.4 \cdot 10^{-5}$ |
| $a^1$ | $b^0$ | $c^1$ | $d^1$ | 100 | $1.4 \cdot 10^{-5}$ |
| $a^1$ | $b^1$ | $c^0$ | $d^0$ | 10 | $1.4 \cdot 10^{-6}$ |
| $a^1$ | $b^1$ | $c^0$ | $d^1$ | 100,000 | 0.014 |
| $a^1$ | $b^1$ | $c^1$ | $d^0$ | 100,000 | 0.014 |
| $a^1$ | $b^1$ | $c^1$ | $d^1$ | 100,000 | 0.014 |

$\approx$ **0.69**



Image source: Koller and Friedman 2009

| $Assignment$ | | | | $Unnormalized$ | $Normalized$ |
|---|---|---|---|---|---|
| $a^0$ | $b^0$ | $c^0$ | $d^0$ | $300,000$ | $0.04$ |
| $a^0$ | $b^0$ | $c^0$ | $d^1$ | $300,000$ | $0.04$ |
| $a^0$ | $b^0$ | $c^1$ | $d^0$ | $300,000$ | $0.04$ |
| $a^0$ | $b^0$ | $c^1$ | $d^1$ | $30$ | $4.1 \cdot 10^{-6}$ |
| $a^0$ | $b^1$ | $c^0$ | $d^0$ | $500$ | $6.9 \cdot 10^{-5}$ |
| $a^0$ | $b^1$ | $c^0$ | $d^1$ | $500$ | $6.9 \cdot 10^{-5}$ |
| $a^0$ | $b^1$ | $c^1$ | $d^0$ | $5,000,000$ | $0.69$ |
| $a^0$ | $b^1$ | $c^1$ | $d^1$ | $500$ | $6.9 \cdot 10^{-5}$ |
| $a^1$ | $b^0$ | $c^0$ | $d^0$ | $100$ | $1.4 \cdot 10^{-5}$ |
| $a^1$ | $b^0$ | $c^0$ | $d^1$ | $1,000,000$ | $0.14$ |
| $a^1$ | $b^0$ | $c^1$ | $d^0$ | $100$ | $1.4 \cdot 10^{-5}$ |
| $a^1$ | $b^0$ | $c^1$ | $d^1$ | $100$ | $1.4 \cdot 10^{-5}$ |
| $a^1$ | $b^1$ | $c^0$ | $d^0$ | $10$ | $1.4 \cdot 10^{-6}$ |
| $a^1$ | $b^1$ | $c^0$ | $d^1$ | $100,000$ | $0.014$ |
| $a^1$ | $b^1$ | $c^1$ | $d^0$ | $100,000$ | $0.014$ |
| $a^1$ | $b^1$ | $c^1$ | $d^1$ | $100,000$ | $0.014$ |

$\approx$ **0.14**



Image source: Koller and Friedman 2009

| Assignment | | | | Unnormalized | Normalized |
|---|---|---|---|---|---|
| $a^0$ | $b^0$ | $c^0$ | $d^0$ | $300,000$ | $0.04$ |
| $a^0$ | $b^0$ | $c^0$ | $d^1$ | $300,000$ | $0.04$ |
| $a^0$ | $b^0$ | $c^1$ | $d^0$ | $300,000$ | $0.04$ |
| $a^0$ | $b^0$ | $c^1$ | $d^1$ | $30$ | $4.1 \cdot 10^{-6}$ |
| $a^0$ | $b^1$ | $c^0$ | $d^0$ | $500$ | $6.9 \cdot 10^{-5}$ |
| $a^0$ | $b^1$ | $c^0$ | $d^1$ | $500$ | $6.9 \cdot 10^{-5}$ |
| $a^0$ | $b^1$ | $c^1$ | $d^0$ | $5,000,000$ | $0.69$ |
| $a^0$ | $b^1$ | $c^1$ | $d^1$ | $500$ | $6.9 \cdot 10^{-5}$ |
| $a^1$ | $b^0$ | $c^0$ | $d^0$ | $100$ | $1.4 \cdot 10^{-5}$ |
| $a^1$ | $b^0$ | $c^0$ | $d^1$ | $1,000,000$ | $0.14$ |
| $a^1$ | $b^0$ | $c^1$ | $d^0$ | $100$ | $1.4 \cdot 10^{-5}$ |
| $a^1$ | $b^0$ | $c^1$ | $d^1$ | $100$ | $1.4 \cdot 10^{-5}$ |
| $a^1$ | $b^1$ | $c^0$ | $d^0$ | $10$ | $1.4 \cdot 10^{-6}$ |
| $a^1$ | $b^1$ | $c^0$ | $d^1$ | $100,000$ | $0.014$ |
| $a^1$ | $b^1$ | $c^1$ | $d^0$ | $100,000$ | $0.014$ |
| $a^1$ | $b^1$ | $c^1$ | $d^1$ | $100,000$ | $0.014$ |

$\approx$ **0.04**

Image source: Koller and Friedman 2009

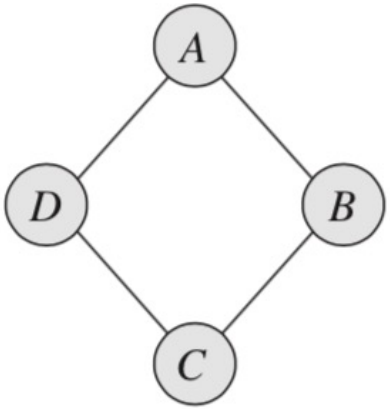
# Compare

Marginal probability $p(A, B)$:

| A | B | P(A,B) |
|---|---|---|
| 0 | 0 | 0.13 |
| 0 | 1 | 0.69 |
| 1 | 0 | 0.14 |
| 1 | 1 | 0.04 |

Conditional probability $p(A|B)$:

| A | B | P(A\|B) |
|---|---|---|
| 0 | 0 | 0.46 |
| 0 | 1 | 0.94 |
| 1 | 0 | 0.54 |
| 1 | 1 | 0.06 |




$$\varphi(A, B)$$

| | | |
|---|---|---|
| $a^0$ | $b^0$ | 30 |
| $a^0$ | $b^1$ | 5 |
| $a^1$ | $b^0$ | 1 |
| $a^1$ | $b^1$ | 10 |

Image source: Koller and Friedman 2009

# Question

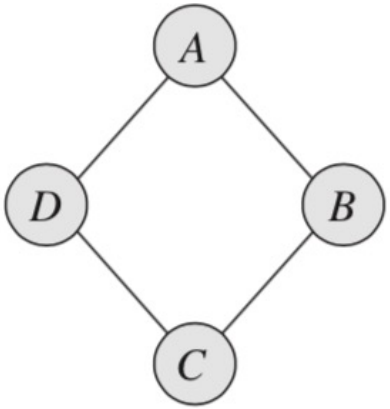
Consider pairwise factor $\varphi(A, B)$. Is the factor proportional to:

A. The marginal probability $p(A, B)$

B. The conditional probability $p(A|B)$

C. All of the above

**D. None of the above**




$\varphi(A, B)$

| | | |
|---|---|---|
| $a^0$ | $b^0$ | 30 |
| $a^0$ | $b^1$ | 5 |
| $a^1$ | $b^0$ | 1 |
| $a^1$ | $b^1$ | 10 |

# UGM Misconceptions

- Some misconceptions about UGMs / MRFs:
  - ✗ Factors always represent marginal/conditional distributions.
  - ✗ Undirected graphs are "richer" (can represent more conditional independencies) compared to DGMs
  - ✗ UGMs specify a unique factorization.

# Parameterization of MRFs

- **Con:** Lose the ability to give local probabilistic interpretation to the functions used to represent the joint probability.

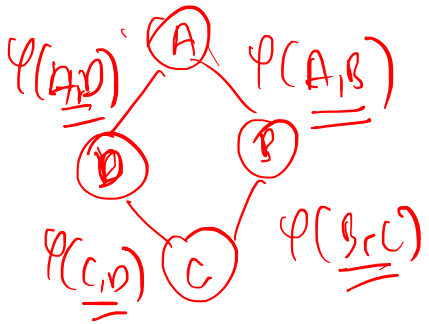
- **Pro:** Retains the ability to model the joint probability as a product of local functions.

$$p(x) = \frac{1}{Z} \prod \varphi_i(x_i)$$

# Gibbs Distribution

- **Definition** (*Gibbs Distribution*) A distribution $P$ is a Gibbs distribution parameterized by a set of factors $\{\varphi_1(\mathcal{C}_1), \varphi_2(\mathcal{C}_2), \ldots, \varphi_m(\mathcal{C}_m)\}$ where

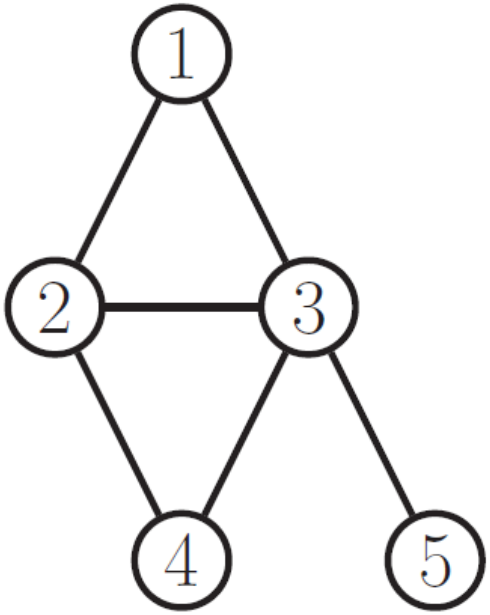
$$p(x_1, x_2, \ldots, x_N) = \frac{1}{Z} \prod_{j=1}^{M} \varphi_j(\mathcal{C}_j)$$




$\varphi(A,B,C,D)$ ?

# Question

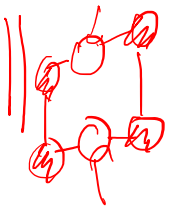
- What should a factorization look like given a graph?
- What should the local functions (factors) be and over which sets of variables?



# Parameterization of MRFs

How do we decide the domain of the local functions?

- Recall two nodes $X_i$ and $X_j$ that are not directly linked in an UGM are conditionally independent given all other nodes.



- Thus it must be possible to obtain a factorization of the joint probability that places $X_i$ and $X_j$ in different factors.

- This implies that we shouldn't have a local function that depends on both $X_i$ and $X_j$.

$$p(x_1, \dots, x_N) \neq \psi_1(x_i, x_j, \dots) \dots \psi_m(\dots)$$

# Parameterization of MRFs

How do we decide the domain of the local functions?

- Our argument thus far suggests that all nodes $X_C$ that belong to a maximal clique $C$ in the UGM appear together in a local function $\psi(x_C)$.
- A clique of a graph is a fully-connected subset of nodes.
- The maximal cliques of a graph are the cliques that cannot be extended to include additional nodes without losing the property of being fully connected.

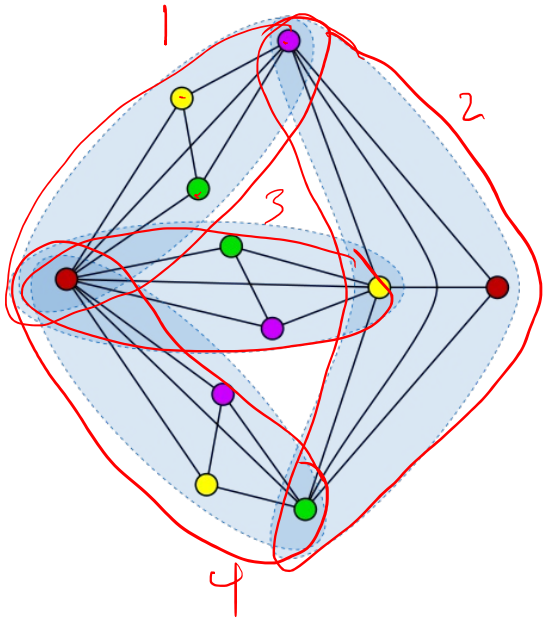
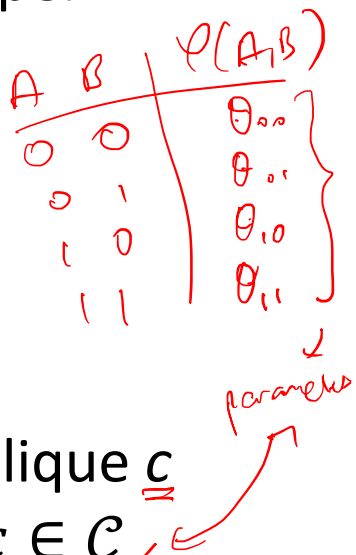



Image source: http://wikivisually.com/wiki/Clique_(graph_theory)

# Hammersley-Clifford

**Theorem 5.1** (*Hammersley-Clifford*) A positive distribution $p(y) > 0$ satisfies the CI properties of an undirected graph $\mathcal{H}$ iff $p$ can be represented as a product of factors, one per maximal clique:

$$p(\mathbf{y}|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \prod_{c \in \mathcal{C}} \psi_c(\mathbf{y}_c|\boldsymbol{\theta}_c)$$




where

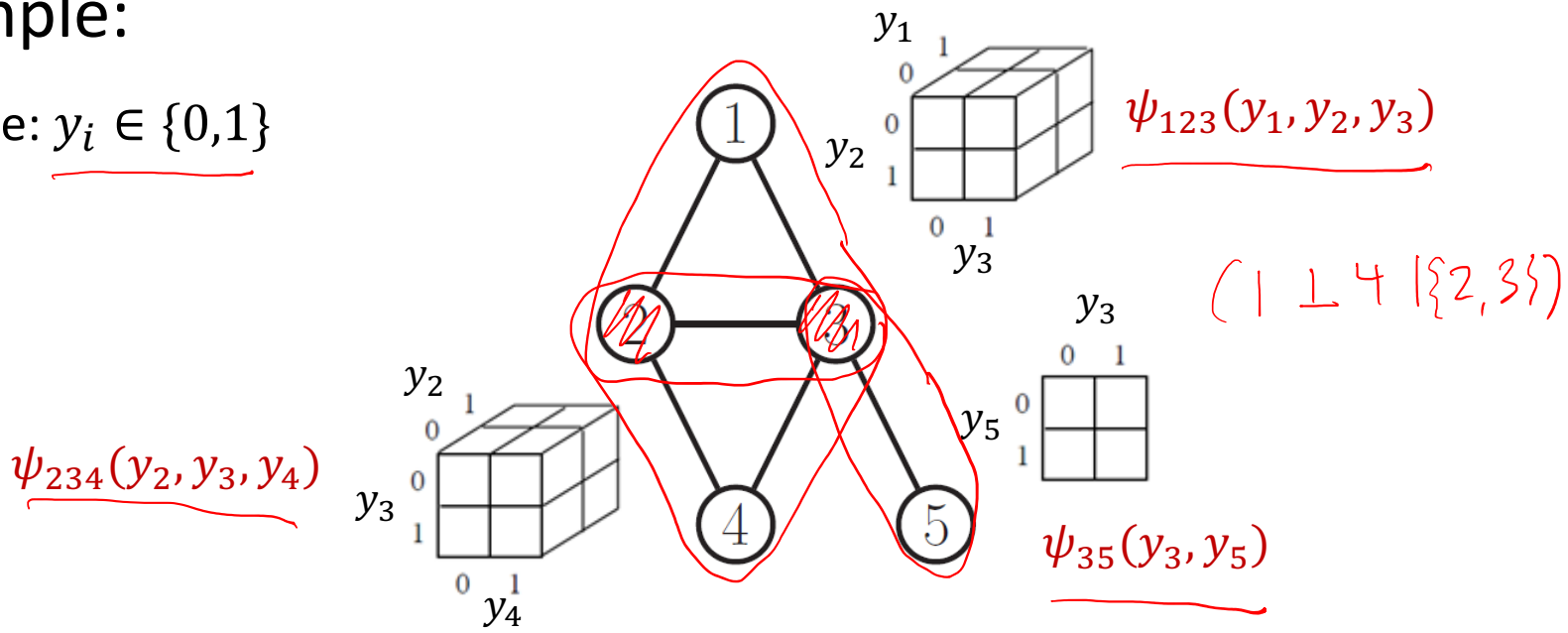
- $\mathcal{C}$ is the set of all the maximal cliques of $\mathcal{G}$
- $\psi_c(.)$ is the factor or potential function of clique $c$
- $\theta$ is the parameter of the factors $\psi_c(.)$ for $c \in \mathcal{C}$
- $Z(\theta)$ is the partition function

$$Z(\boldsymbol{\theta}) \triangleq \sum \prod_{c \in \mathcal{C}} \psi_c(\mathbf{y}_c|\boldsymbol{\theta}_c)$$

# Parameterization of MRFs

## Example:

Assume: $y_i \in \{0,1\}$



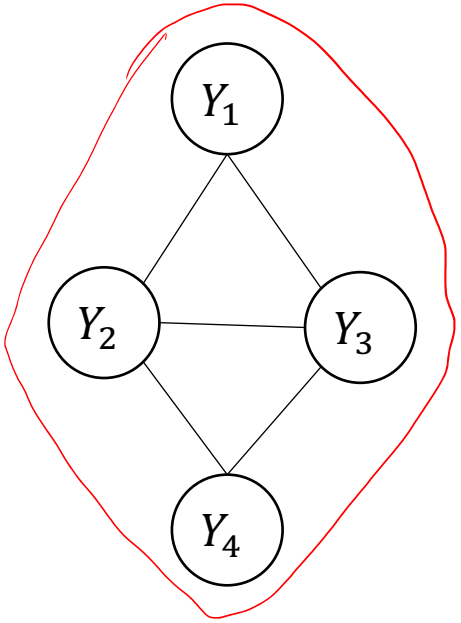

$$p(\mathbf{y}|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \psi_{123}(y_1, y_2, y_3)\psi_{234}(y_2, y_3, y_4)\psi_{35}(y_3, y_5)$$

where $$Z = \sum_{\mathbf{y}} \psi_{123}(y_1, y_2, y_3)\psi_{234}(y_2, y_3, y_4)\psi_{35}(y_3, y_5)$$

"An introduction to probabilistic graphical models", Michael I. Jordan, 2002

# Question

**H:**




**Which factorization is consistent with H?**

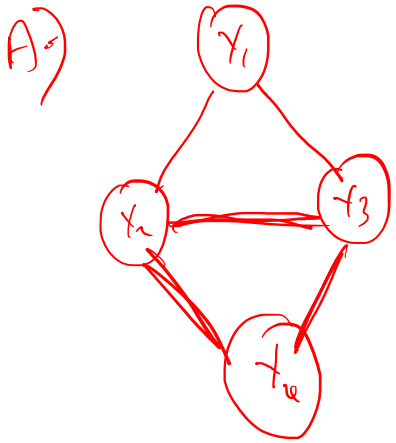
A. $\varphi(y_1, y_2, y_3)\varphi(y_2, y_3, y_4)$

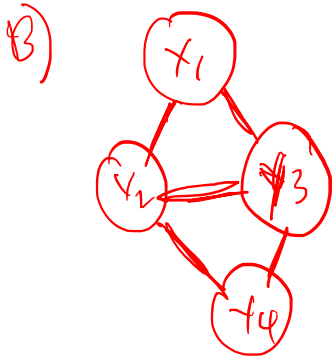
B. $\varphi(y_1, y_2)\varphi(y_1, y_3)\varphi(y_2, y_3)\varphi(y_2, y_4)\varphi(y_3, y_4)$

C. $\varphi(y_1, y_2, y_3)\varphi(y_2, y_4)\varphi(y_3, y_4)$

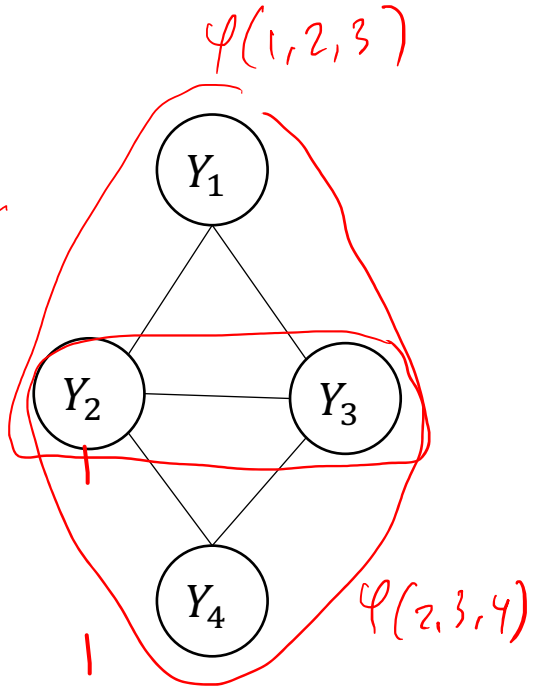
D. All of the above

# Question





**H:**




**Which factorization is consistent with H?**

A. $\varphi(y_1, y_2, y_3)\varphi(y_2, y_3, y_4)$

B. $\varphi(y_1, y_2)\varphi(y_1, y_3)\varphi(y_2, y_3)\varphi(y_2, y_4)\varphi(y_3, y_4)$

C. $\varphi(y_1, y_2, y_3)\varphi(y_2, y_4)\varphi(y_3, y_4)$

**D. All of the above**

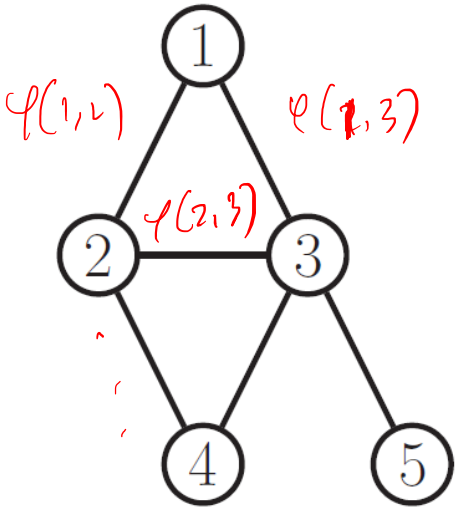
# UGM Misconceptions

- Some misconceptions about UGMs / MRFs:
  - ✗ Factors always represent marginal/conditional distributions.
  - ✗ Undirected graphs are "richer" (can represent more conditional independencies) compared to DGMs
  - ✗ UGMs specify a unique factorization.

# Parameterization of MRFs

- We are free to restrict the parameterization to the edges of the graph, rather than the maximal cliques.
- This is called a pairwise MRF.
- This form is widely used due to its simplicity, although it is not as general.

**Why not as general?**




Example 1:

$$p(\mathbf{y}|\boldsymbol{\theta}) \propto \psi_{12}(y_1, y_2)\psi_{13}(y_1, y_3)\psi_{23}(y_2, y_3)\psi_{24}(y_2, y_4)\psi_{34}(y_3, y_4)\psi_{35}(y_3, y_5)$$
$$\propto \prod_{s \sim t} \psi_{st}(y_s, y_t)$$

"An introduction to probabilistic graphical models", Michael I. Jordan, 2002

# Representing Potential Functions

- Potentials represent the relative "compatibility" between the different assignments to the random variables.
- A general approach is to define the log potentials as a linear function of the parameters:

$$\log \psi_c(\mathbf{y}_c) \triangleq \boldsymbol{\phi}_c(\mathbf{y}_c)^T \boldsymbol{\theta}_c$$

features  coefficients/parameters

- $\phi_c(y_c)$ is a feature vector derived from the values of the variables $y_c$.

# Representing Potential Functions

**Example:**

Consider a pairwise MRF, where we associate a feature vector of length $K^2$ for each edge as follows:

$$\phi_{st}(y_s, y_t) = [\ldots, \mathbb{I}(y_s = j, y_t = k), \ldots]$$

Indicator function that returns 1 when conditions are true, 0 otherwise

If we have a weight for each feature, we can convert this into a $K \times K$ potential function (tabular) as follows:

$$\psi_{st}(y_s = j, y_t = k) = \exp([\boldsymbol{\theta}_{st}^T \boldsymbol{\phi}_{st}]_{jk}) = \exp(\theta_{st}(j,k))$$

$K^2 \times 1$

# Example: Simple Log-Linear Model


$\psi(y_1, y_2)$

| $y_1$ | $y_2$ | $\psi(y_1, y_2)$ |
|---|---|---|
| 0 | 0 | $a_{00}$ |
| 0 | 1 | $a_{01}$ |
| 1 | 0 | $a_{10}$ |
| 1 | 1 | $a_{11}$ |

Define features:

$$\phi_{12}^{00} = \mathbb{I}[y_1 = 0, y_2 = 0]$$
$$\phi_{12}^{01} = \mathbb{I}[y_1 = 0, y_2 = 1]$$
$$\phi_{12}^{10} = \mathbb{I}[y_1 = 1, y_2 = 0]$$
$$\phi_{12}^{11} = \mathbb{I}[y_1 = 1, y_2 = 1]$$

Then,

$$\psi(y_1, y_2) = \exp\left(\sum_{kl} \theta_{kl} \phi_{ij}^{kl}(y_i, y_j)\right)$$

where $\theta_{kl} = \log a_{kl}$

$$p(y_1, y_2) = \frac{1}{Z(\boldsymbol{\theta})} \exp(\boldsymbol{\theta}^\top \boldsymbol{\phi}(\boldsymbol{y}))$$

# Representing Potential Functions

- The resulting log probability has the form:

$$\log p(\mathbf{y}|\boldsymbol{\theta}) = \sum_c \boldsymbol{\phi}_c(\mathbf{y}_c)^T \boldsymbol{\theta}_c - \log Z(\theta)$$

- This is also known as a maximum entropy or a log-linear model.

$$p(\mathbf{y}|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \exp\left(\sum_c \boldsymbol{\theta}_c^T \boldsymbol{\phi}_c(\mathbf{y})\right)$$

# Relationship to ExpFam

- **Claim:** Every finite MRF is an exponential family.
- **Recall from Lecture 2:** An exponential family (ExpFam) is a set of probability distributions $\{p_\theta : \theta \in \Theta\}$ with the form

$$p_\theta(x) = \frac{h(x)\exp[\eta(\theta)^\top s(x)]}{Z(\theta)}$$

- where:
  - $\theta \in \mathbb{R}^k, x \in \mathbb{R}^d$
  - Natural parameters: $\eta(\theta): \Theta \to \mathbb{R}^m$
  - Sufficient statistics: $s(x): \mathbb{R}^d \to \mathbb{R}^m$
  - Base Measure (Support and scaling): $h(x): \mathbb{R}^d \to [0, \infty)$
  - Partition function: $\mathrm{Z}(\theta): \Theta \to [0, \infty)$

NUS National University of Singapore | School of Computing

# MRF Examples

# Representing Potential Functions

- The resulting log probability has the form:

$$\log p(\mathbf{y}|\boldsymbol{\theta}) = \sum_c \boldsymbol{\phi}_c(\mathbf{y}_c)^T \boldsymbol{\theta}_c - \log Z(\theta)$$

- This is also known as a maximum entropy or a log-linear model.

$$p(\mathbf{y}|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \exp\left(\sum_c \boldsymbol{\theta}_c^T \boldsymbol{\phi}_c(\mathbf{y})\right)$$

weight feature vectors.

# Representing Potential Functions

- The resulting log probability has the form:

$$\log p(\mathbf{y}|\boldsymbol{\theta}) = \sum_c \boldsymbol{\phi}_c(\mathbf{y}_c)^T \boldsymbol{\theta}_c - \log Z(\theta)$$

- This is also known as a maximum entropy or a log-linear model.

$$p(\mathbf{y}|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \exp\left(\sum_c \boldsymbol{\theta}_c^T \boldsymbol{\phi}_c(\mathbf{y})\right)$$

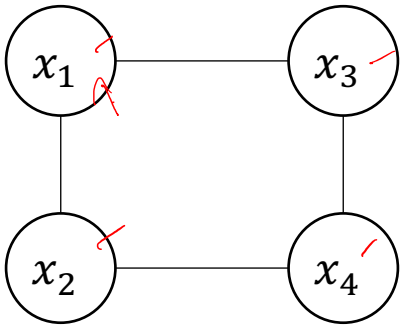
- Can also specify

$$p(\mathbf{y}|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \exp(-\sum_c E(\mathbf{y}_c|\boldsymbol{\theta}_c))$$

  - where $E(y_c|\theta_c)$ is the energy associated with the variables in clique *c*.

# Ising Model

Each $x_{ij} \in \{-1, +1\}$



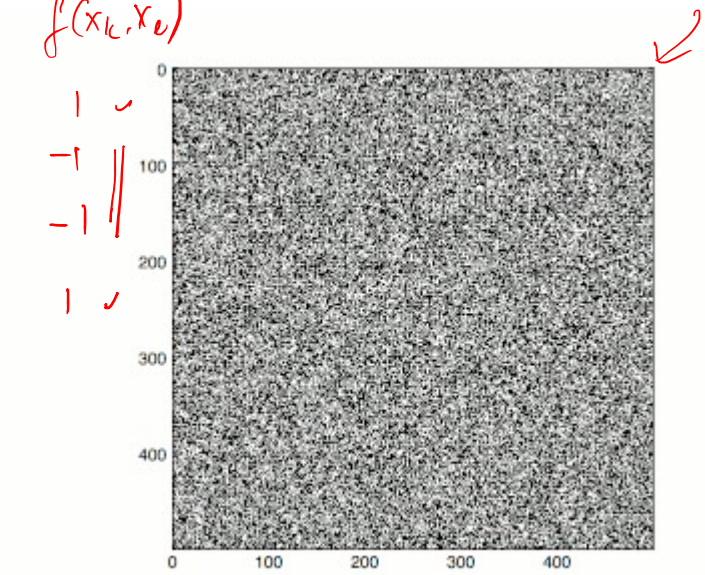

$$p(x|\theta) = \frac{1}{Z(\theta)} \exp\left(-\sum_c E(x_c|\theta_c)\right)$$

where

$$\sum_c E(x_c|\theta_c) = -\sum_{(k,l)} \theta_{kl} x_k x_l - \sum_m \vartheta_m x_m$$

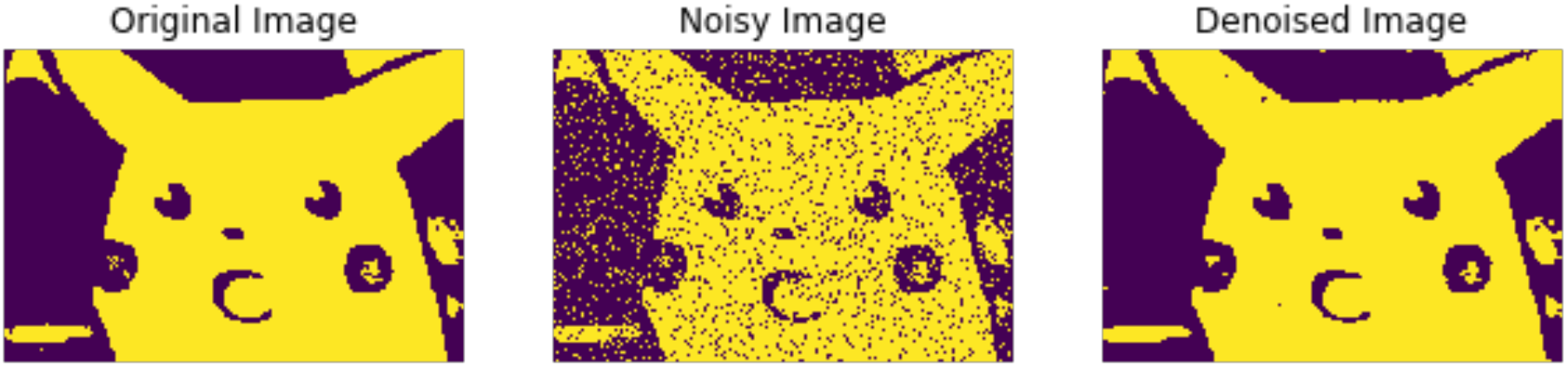
Pairwise potential

Unary potential

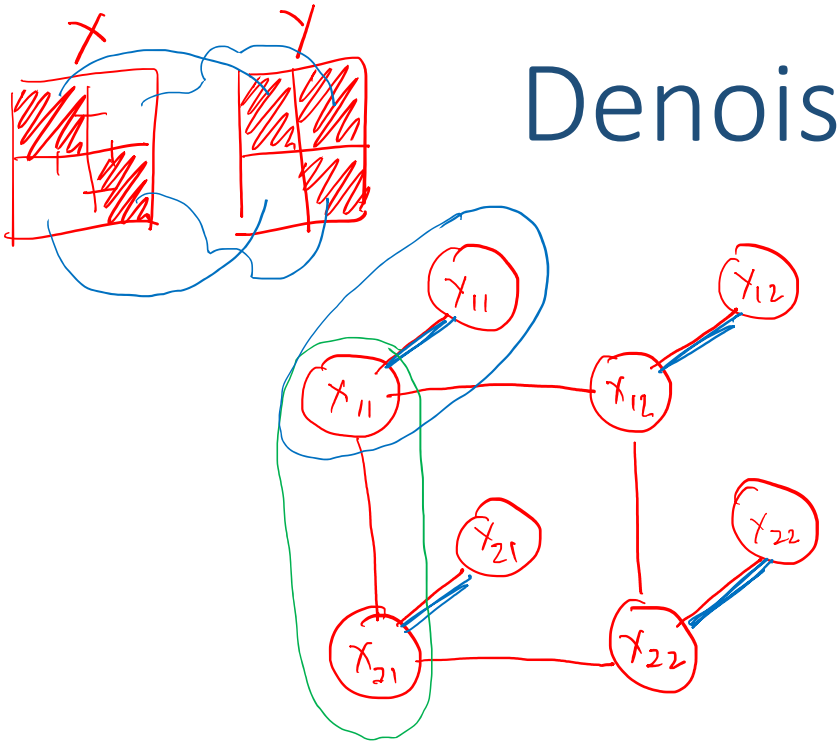



An Ising system on a two dimensional square lattice (500x500) with temperature $T = 0.1$ starting from a random configuration. (source: Wikipedia: https://en.wikipedia.org/wiki/Ising_model)

# Denoising



# Denoising MRF




$$p(x|\theta) = \frac{1}{Z(\theta)} \exp\left(-\sum_c E(x_c|\theta_c)\right)$$

low energy when agreement happens.

neighbor of $x_{ij}$

$$\psi(x_{ij}, x'_{ij}) = -\beta x_{ij} x'_{ij}$$

$\beta > 0$

$\alpha = 2$

$\psi(x_{ij}, y_{ij}) = ?$

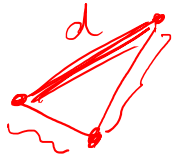
| $x_{ij}$ | $x'_{ij}$ | $x_{ij}x'_{ij}$ | $-\alpha x_{ij}x'_{ij}$ | |
|---|---|---|---|---|
| -1 | -1 | +1 | -2 | // low |
| -1 | +1 | -1 | +2 | // |
| +1 | -1 | -1 | +2 | // |
| +1 | +1 | +1 | -2 | // low |

# Metric MRFs

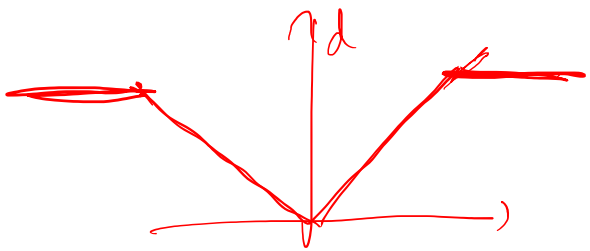
- All $X_i$ take values in some label space $V$ e.g.,
  - $V = \{-1, 1\}$
  - $V = \{\text{dog}, \text{cat}, \text{monkey}\}$
  - $V = \{0,1,2,3,4,5\}$
- **Key idea:** If $X_i$ and $X_j$ connected, we want them to take on "similar" values.

# Metric MRFs: Distance Function

- Define a distance function: $d: V \times V \rightarrow R^+$
  - **Reflexivity:** $\forall v \in V,\ d(v, v) = 0$
  - **Symmetry:** $\forall v_i, v_j \in V\ ,\ d(v_i, v_j) = d(v_j, v_i)$
  - **Triangle inequality:** $\forall v_i, v_j, v_k \in V\ , d(v_i, v_j) \leq d(v_i, v_k) + d(v_k, v_j)$




- Examples:
  - **Euclidean metric:** $d(v_i, v_j) = \left(v_i^2 + v_j^2\right)^{1/2}$
  - **Discrete metric:** $d(v_i, v_j) = 0$ if $v_i = v_j$ (and 1 otherwise)
  - **Taxicab metric:** $d(v_i, v_j) = |v_i - v_j|$
  - Truncated linear penalty
  - Hamming distance
  - Wasserstein metric
  - Etc.



# Metric MRFs: Distance as a Feature

- Given a distance function $d(x_i, x_j)$
- Define:

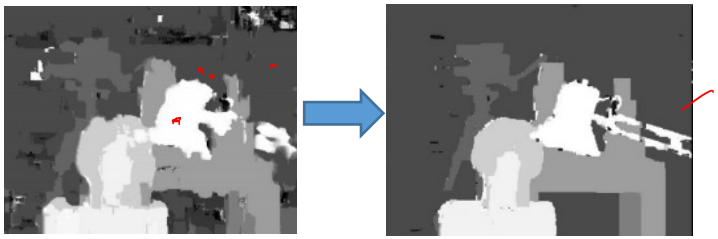
$$\psi_{ij}(x_i, x_j) = d(x_i, x_j)$$

- Hence:

$$\exp\left(-\theta_{ij}\psi_{ij}(x_i, x_j)\right)$$

where $\theta_{ij} > 0$

Depth Map from Stereo



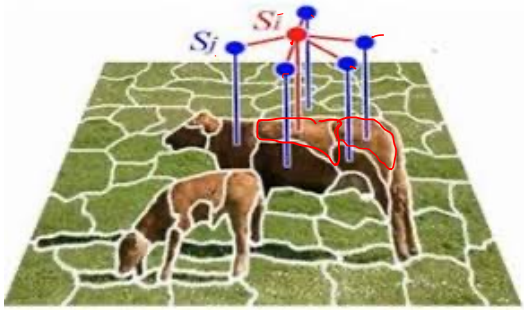



Image Source: https://cedar.buffalo.edu/~srihari/CSE574/Chap8/Ch8-PGM-Undirected/9.5-MRFinCV.pdf

# Language Model

**Example:**

- Suppose we are interested in making a probabilistic model of English spelling.
- We need higher order factors to capture certain letter combinations occur together quite frequently (e.g. “ing”).
- Suppose we limit ourselves to letter trigrams, a tabular potential still has $26^3 = 17{,}576$ parameters in it.
- However, most of these triples will never occur.

# Language Model

**Example:**

- An alternative approach is to define indicator functions that look for certain "special" triples, such as "ing", "qu-", etc.

- Then we can define the potential on each trigram as follows:

$$\psi(y_{t-1}, y_t, y_{t+1}) = \exp(\sum_k \theta_k \phi_k(y_{t-1}, y_t, y_{t+1}))$$

- $k$ indexes the different features, corresponding to "ing", "qu-", etc., and $\phi_k$ is the corresponding binary feature function.

# Conditional Random Fields

- A CRF or discriminative random field, is just a version of an MRF where all the clique potentials are conditioned on input features X:

$$p(\mathbf{y}|\mathbf{x},\mathbf{w}) = \frac{1}{Z(\mathbf{x},\mathbf{w})}\prod_c \psi_c(\mathbf{y}_c|\mathbf{x},\mathbf{w})$$

- We will usually assume a log-linear representation of the potentials:

$$\psi_c(\mathbf{y}_c|\mathbf{x},\mathbf{w}) = \exp(\mathbf{w}_c^T \boldsymbol{\phi}(\mathbf{x},\mathbf{y}_c))$$

- where $\phi(\mathrm{x}, y_c)$ is a feature vector derived from the global inputs $\mathrm{X}$ and the local set of labels $Y_c$.

# Discriminative Vs Generative Models

- Generative models: Approaches that explicitly or implicitly model the distribution of inputs and outputs.
- Sampling from the distribution it is possible to generate synthetic data points in the input space.

**Likelihood**: $p(\mathbf{x}|\mathcal{C}_k)$

$p(x,y)$ likelihood · prior.

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

- Discriminative models: Approaches that model the posterior probabilities directly.

**Posterior**: $p(\mathcal{C}_k|\mathbf{x})$

$p(y|x) = \mathcal{N}(y|w^T x, \sigma_n^2)$

$= p(y|x,w) = \frac{1}{1+\exp(x^T w)}$

# Conditional Random Fields

**Example**: models for sequential data

Hidden Markov model:

$$p(\mathbf{x}, \mathbf{y}|\mathbf{w}) = \prod_{t=1}^{T} p(y_t|y_{t-1}, \mathbf{w}) p(\mathbf{x}_t|y_t, \mathbf{w})$$

Likelihood, i.e. generative

Chain structure MRF:

$$p(\mathrm{x}, \mathrm{y} \mid \mathrm{w}) \propto \prod_{t=1}^{T} p(x_t \mid y_t, \mathrm{w}) \prod_{t=1}^{T-1} \psi(y_t, y_{t+1} \mid \mathrm{w})$$

Likelihood, i.e. generative

Chain structure CRF:

$$p(\mathbf{y}|\mathbf{x}, \mathbf{w}) = \frac{1}{Z(\mathbf{x}, \mathbf{w})} \prod_{t=1}^{T} \psi(y_t|\mathbf{x}, \mathbf{w}) \prod_{t=1}^{T-1} \psi(y_t, y_{t+1}|\mathbf{x}, \mathbf{w})$$

Posteriors, i.e. discriminative

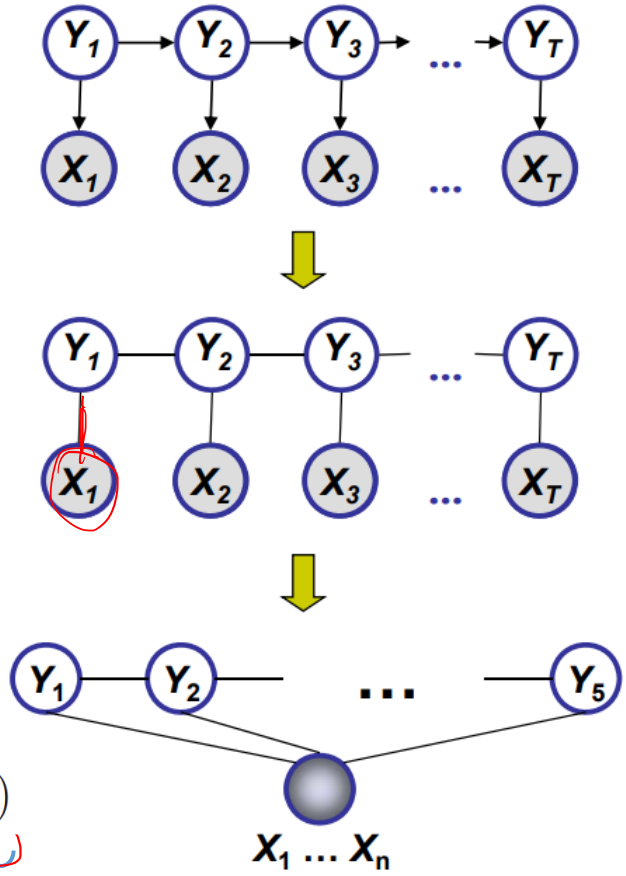
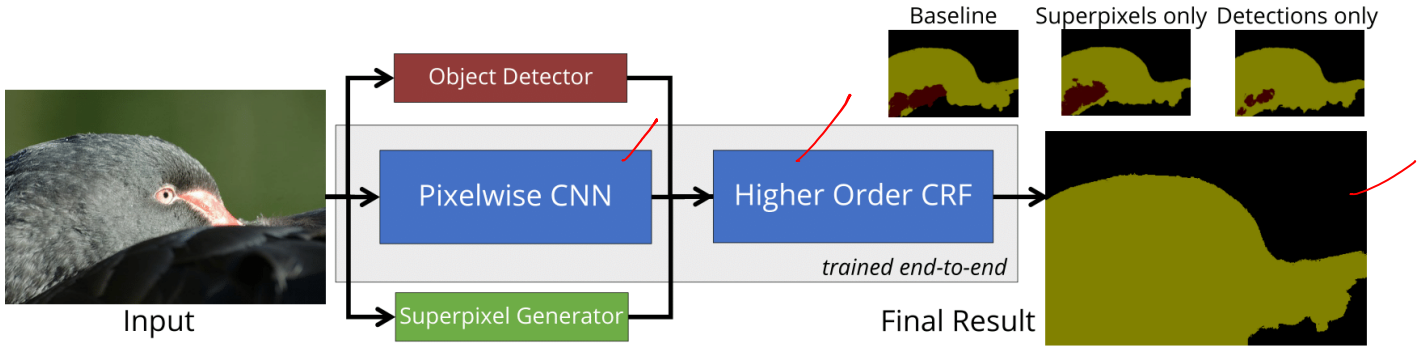
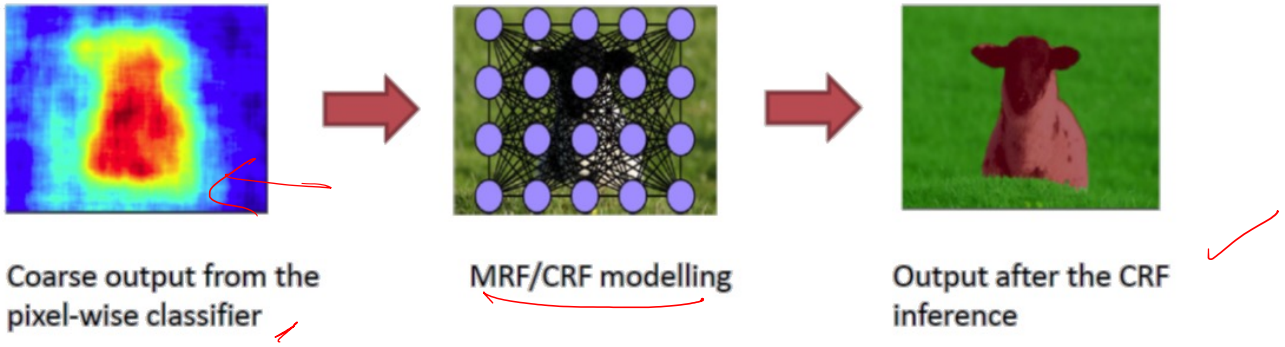



Image source: http://www.cs.cmu.edu/~epxing/Class/10708/lectures/lecture3-MRFrepresentation.pdf, Eric Xing

# Image Semantic Segmentation




Arnab, Anurag, et al. "Higher order conditional random fields in deep neural networks." *European Conference on Computer Vision*. Springer, Cham, 2016.




Zheng, Shuai, et al. "Conditional random fields as recurrent neural networks." *Proceedings of the IEEE international conference on computer vision*. 2015.

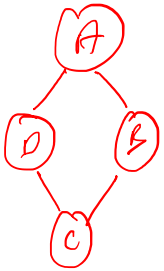
NUS National University of Singapore | School of Computing

# MRF Theory

*Soundness and Completeness*

# Recall from Lecture 3:
# General Questions about Bayes Nets

- Is the Bayesian Network correct/sound?
  - Does a conditional independence identified by d-separation always exist in the distribution? **Yes.** ☺
- Is the Bayesian Network complete?
  - If a conditional independence exists in the distribution, can it always be detected by d-separation? **Almost Yes.** 😐
- How expressive are Bayesian Networks as a modeling language?
  - Can they exactly represent all conditional independencies for a given distribution? **No.** ☹




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

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

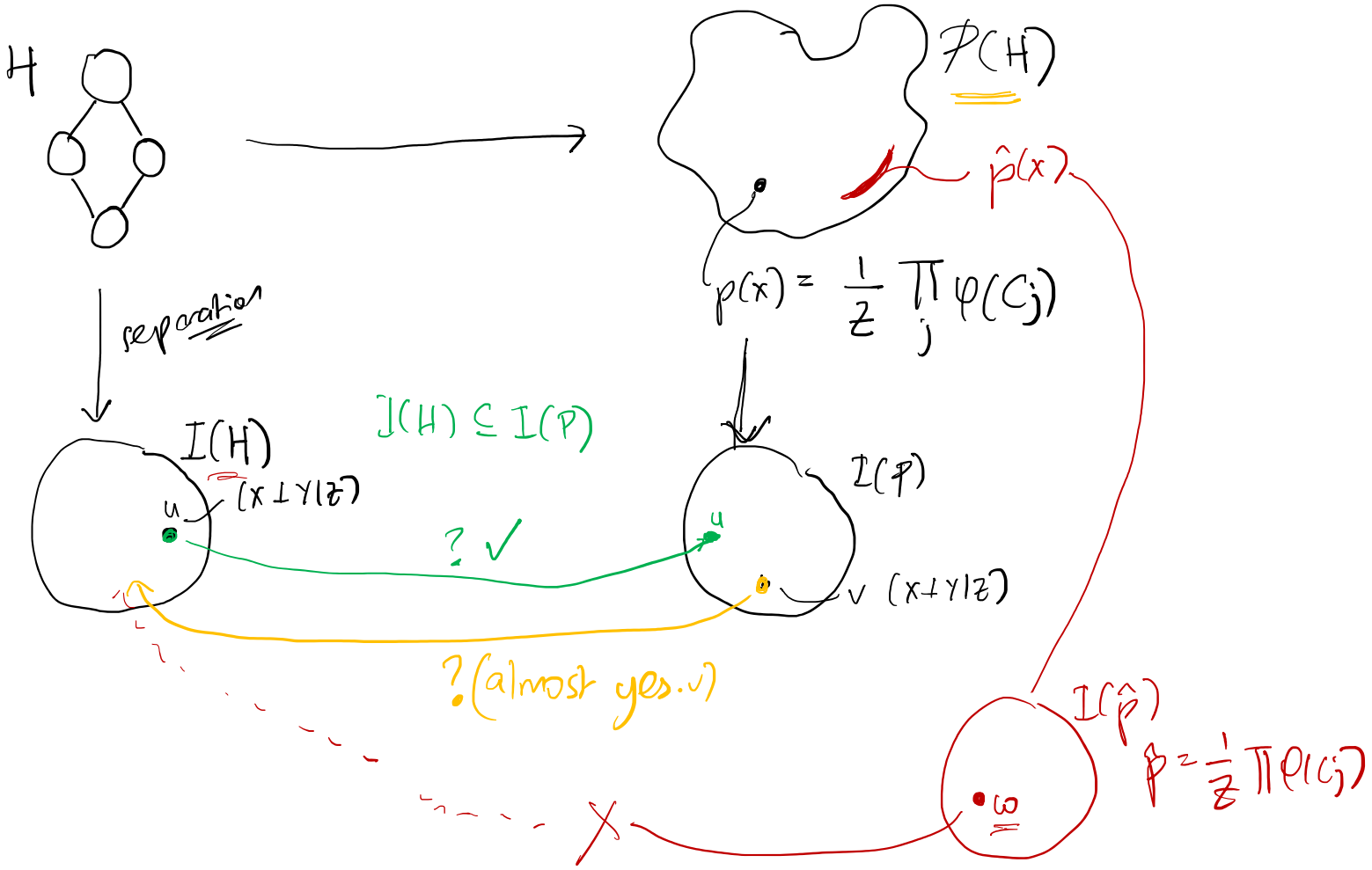
# Questions about MRFs

- Is the MRF correct/sound?
  - Does a conditional independence identified by separation always exist in the distribution?
- Is the MRF complete?
  - If a conditional independence exists in the distribution, can it always be detected by separation?
- How expressive are MRF as a modeling language?
  - Can they exactly represent all conditional independencies for a given distribution?

# Recall: Independence-Maps (I-Maps)

- **Definition** (*Independence set*) Let $P$ be a distribution over $\mathcal{X}$. Define $\mathcal{I}(P)$ as the set of independence assertions of the form $(X \perp Y \mid Z)$ that hold in $P$.

- **Definition** (*Independence map*) Let $G$ be associated with independence assertions $\mathcal{I}(G)$. $G$ is an I-map for $P$ if $\mathcal{I}(G) \subseteq \mathcal{I}(P)$


H
P(H)
p̂(x)
p(x) = 1/Z ∏_j φ(C_j)
separation
I(H)
(z ⊥ y | x)
I(H) ⊆ I(P)
I(P)
u
u
v
(z ⊥ y | x)
? ✓
? (almost yes..)
I(p̂)
p̂ = 1/Z ∏ φ(C_j)
ω

# Soundness and Completeness in UGMs

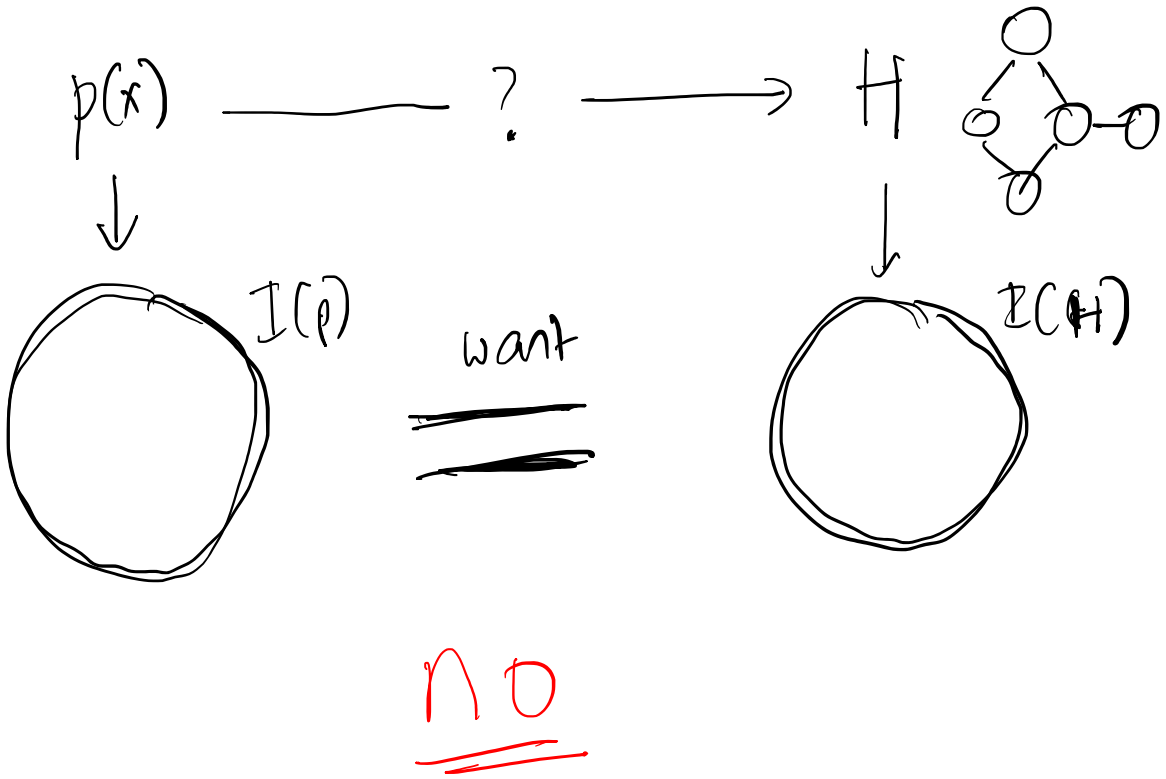
- **Theorem 5.2** (*Soundness*) If $P$ is a Gibbs distribution over $H$, then $H$ is an I-map for $P$
  - Hammersley-Clifford states that iff $H$ is an I-map for $P$, then $P$ is a Gibbs distribution over $H$ (for positive distributions)

- **Theorem 5.3** ("Weak" Completeness) If $X$ and $Y$ are not separated in $H$, then there is some distribution $P$ that factorizes over $H$ where $X$ and $Y$ are dependent

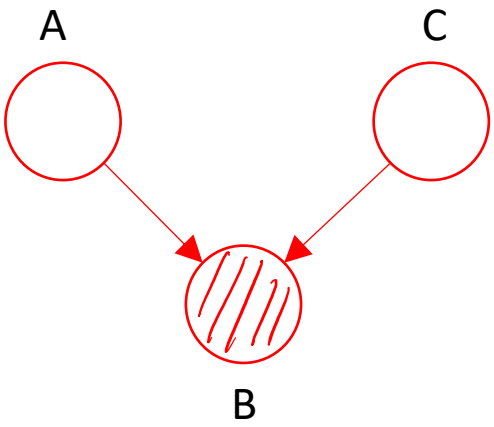
# Expressiveness: Can *H* represent all the independencies for a given *P*?

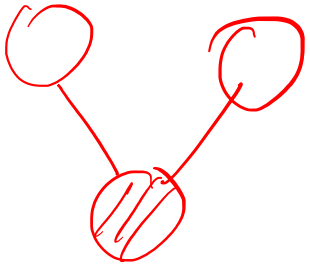
- Can we find a MRFs that represents all the conditional independencies in a given probability distribution $P$?
- **Definition** (*Perfect Map*) A graph $G$ is a perfect map for a probability distribution $P$ if $\mathcal{I}(G) = \mathcal{I}(P)$.

Does every distribution have a perfect map?


p(x)
?
H
I(p)
want
I(H)
NO

# Counter-example



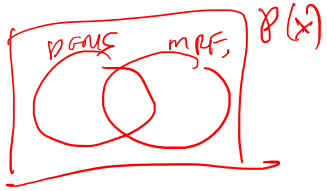
# Questions about MRFs

- Is the MRF correct/sound?
  - Does a conditional independence identified by separation always exist in the distribution? **Yes.** ☺
- Is the MRF complete?
  - If a conditional independence exists in the distribution, can it always be detected by separation? **Almost Yes.** 😐
- How expressive are MRF as a modeling language?
  - Can they exactly represent all conditional independencies for a given distribution?

**No.** ☹

# UGM Misconceptions

- Some misconceptions about UGMs / MRFs:
  - ✗ Factors always represent marginal/conditional distributions.
  - ✗ Undirected graphs are "richer" (can represent more conditional independencies) compared to DGMs
  - ✗ UGMs specify a unique factorization.
- These are incorrect!
- We will learn the right way to interpret UGMs.
- Just like DGMs, UGMs encode a set of conditional independence assertions.



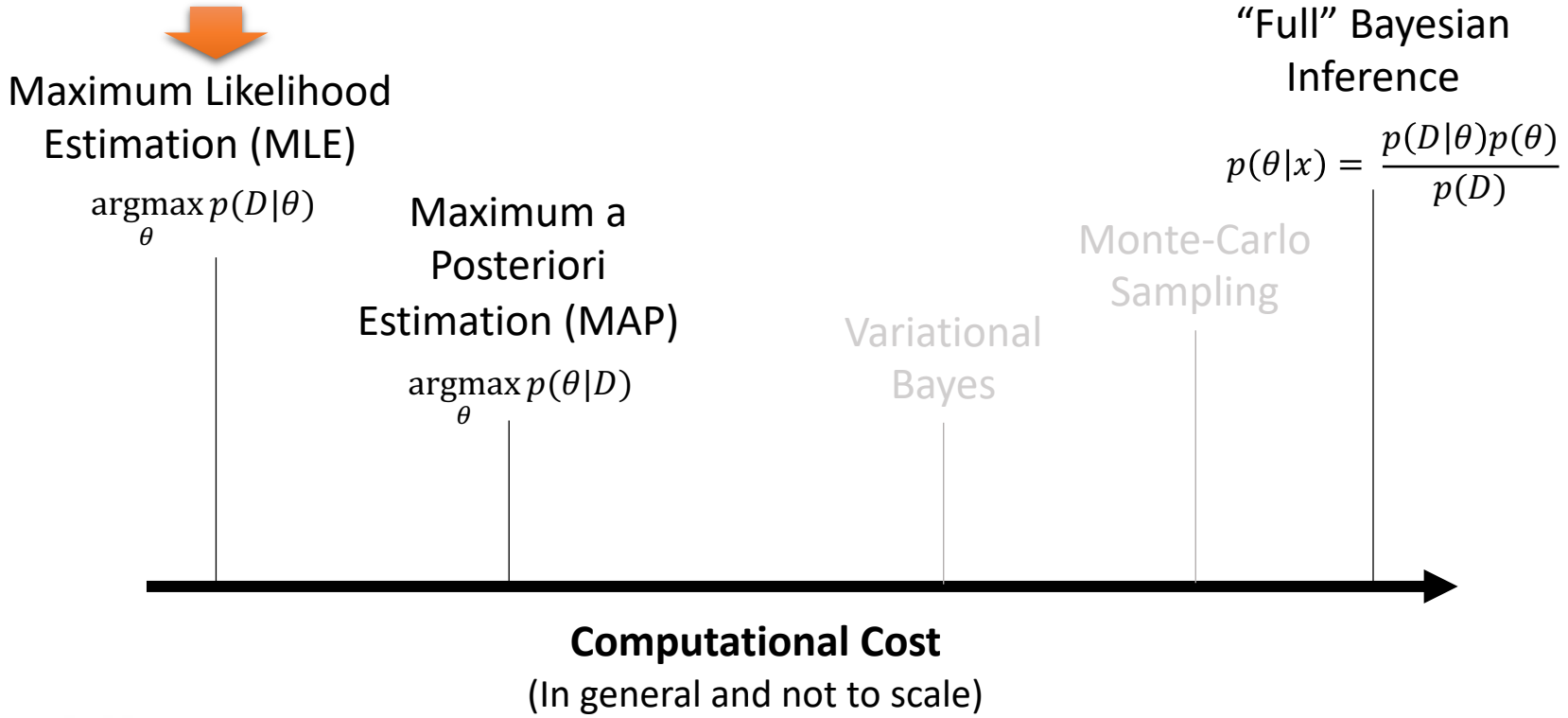
NUS National University of Singapore | School of Computing

# Parameter Learning

*Stochastic Maximum Likelihood for MRFs*

# From Lecture 2: Learning Parameters

- Common approaches to learn the unknown parameters $\theta$ from a set of given data $\mathcal{D} = \{x[1], \ldots, x[N]\}$:



# Parameter Learning: UGM (MRF)

- Consider a Markov Random Field (MRF) in log-linear form, where $c$ indexes the cliques:

$$p(\mathbf{y}|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \exp\left(\sum_c \boldsymbol{\theta}_c^T \boldsymbol{\phi}_c(\mathbf{y})\right)$$

- The scaled log-likelihood is given by:

$$\ell(\boldsymbol{\theta}) \triangleq \frac{1}{N} \sum_{i}^{N} \log p(\mathbf{y}_i|\boldsymbol{\theta}) = \frac{1}{N} \sum_{i}^{N} \left[\sum_c \boldsymbol{\theta}_c^T \boldsymbol{\phi}_c(\mathbf{y}_i) - \log Z(\boldsymbol{\theta})\right]$$

# From Lecture 2: Natural/Canonical form

- An exponential family is in its natural (canonical) form if it is parameterized by its natural parameters:

$$p_\eta(x) = p(x|\eta) = \frac{h(x)\exp[\eta^\top s(x)]}{Z(\eta)}$$

Compare against MRF log-linear form

$$p(\mathbf{y}|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})}\exp\left(\sum_c \boldsymbol{\theta}_c^T \boldsymbol{\phi}_c(\mathbf{y})\right)$$

# Parameter Learning: UGM (MRF)

- Discrete MRFs are in the exponential family and this function is convex in $\theta$.
- So it has a unique global maximum, which we can find using gradient-based optimizers.
- In particular, the derivative for the weights of a particular clique $c$ is given by:

$$\frac{\partial \ell}{\partial \boldsymbol{\theta}_c} = \frac{1}{N} \sum_i^N \left[ \boldsymbol{\phi}_c(\mathbf{y}_i) - \frac{\partial}{\partial \boldsymbol{\theta}_c} \log Z(\boldsymbol{\theta}) \right]$$

# From Lecture 2: Moments of Sufficient Statistics

- For any exponential family distribution:

$$\mathbb{E}[s(x)] = \nabla \log \mathrm{Z}(\eta)$$

- Higher order moments of $s(x)$ given by higher order derivatives.
- If $s(x) = x$ (natural exponential family), we can find moments of $x$ simply by differentiation!

# Parameter Learning: UGM (MRF)

- The derivative of the log partition function w.r.t. $\theta_c$ is the expectation of the $c^{th}$ feature under the model:

$$\frac{\partial \log Z(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_c} = \mathbb{E}\left[\phi_c(\mathbf{y})|\boldsymbol{\theta}\right] = \sum_{\mathbf{y}} \phi_c(\mathbf{y}) p(\mathbf{y}|\boldsymbol{\theta})$$

feature vectors.

**Proof Sketch:**

$$\frac{\partial \log Z(\theta)}{\partial \theta_c} = \frac{1}{Z(\theta)} \frac{\partial Z(\theta)}{\partial \theta_c}, \quad \text{where} \quad Z(\theta) = \sum_y \exp\left(\sum_c \theta_c^T \phi_c(y)\right)$$

$$\Rightarrow \frac{\partial Z(\theta)}{\partial \theta_c} = \sum_y \exp\left(\sum_c \theta_c^T \phi_c(y)\right) \phi_c(y)$$

$$\Rightarrow \frac{\partial \log Z(\theta)}{\partial \theta_c} = \frac{1}{Z(\theta)} \sum_y \phi_c(y) \exp\left(\sum_c \theta_c^T \phi_c(y)\right)$$

$$= \sum_y \phi_c(y) \underbrace{\frac{1}{Z(\theta)} \exp\left(\sum_c \theta_c^T \phi_c(y)\right)}_{p(y|\theta)} = \sum_y \phi_c(y)\ p(y|\theta)$$

# Parameter Learning: UGM (MRF)

- Hence the gradient of the log-likelihood is:

$$\frac{\partial \ell}{\partial \boldsymbol{\theta}_c} = \left[\frac{1}{N}\sum_i^N \boldsymbol{\phi}_c(\mathbf{y}_i)\right] - \mathbb{E}\left[\boldsymbol{\phi}_c(\mathbf{y})\right]$$

Clamped term

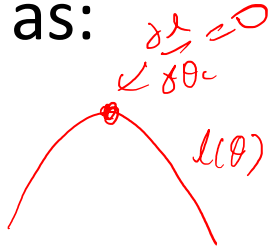
Unclamped/contrastive term

- Clamped term: $y$ is fixed to its observed values.
- Unclamped/contrastive term: $y$ is a free variable.
- Unclamped term requires inference in the model, once per gradient step, and this makes UGM learning much slower than DGM.

# Parameter Learning: UGM (MRF)

- Gradient of the log-likelihood can be rewritten as:

$$\frac{\partial l}{\partial \theta_c} = E_{p_{emp}}[\phi_c(y)] - E_{p(y|\theta)}[\phi_c(y)]$$




- $E_{p_{emp}}[\phi_c(y)] = \frac{1}{N}\sum_{n=1}^{N} \phi_c(y_i)$: Expected feature vector according to the empirical distribution.

- $E_{p(y|\theta)}[\phi_c(y)]$: Expected feature vector according to the model's distribution.

# Parameter Learning: UGM (MRF)

- At the optimum, the gradient will be zero:

$$E_{p_{emp}}[\phi_c(y)] - E_{p(y|\theta)}[\phi_c(y)] = 0$$

- **Problem**: $E_{p(y|\theta)}[\phi_c(y)] = \sum_y \phi_c(y)\ p(y|\theta)$ cannot be evaluated in closed-form in terms of the unknown parameters $\theta$

# Parameter Learning: UGM (MRF)

**Solution**:

Use gradient-based optimizers!

- However, the gradient requires inference:

Requires sum over all states of y, which is intractable

$$\frac{\partial l}{\partial \theta_c} = E_{p_{emp}}[\phi_c(y)] - E_{p(y|\theta)}[\phi_c(y)]$$

- Gradient is intractable, hence learning also becomes intractable.
- We can combine approximate inference with gradient-based learning.

# Stochastic Maximum Likelihood

- This is a stochastic gradient descent method.
- We iteratively updates the parameter $\theta_{k+1}$ at the $k$ step using the parameter and:

$$\theta_{k+1} \leftarrow \theta_k - \eta g_k$$

- $\eta$ is the step size, or learning rate.
- $g_k \approx \frac{\partial l}{\partial \theta_c}$ is the gradient approximated with Markov Chain Monte Carlo (MCMC), i.e. sampling.

# Stochastic Maximum Likelihood

**Algorithm :** Stochastic maximum likelihood for fitting an MRF

Initialize weights $\boldsymbol{\theta}$ randomly;
$k = 0$, $\eta = 1$ ;
**for** *each epoch* **do**
  **for** *each minibatch of size* $B$ **do**  // split the observed data $\mathrm{y}_i, \forall i = 1 \ldots N$ into sets of size $B$
    **for** *each sample* $s = 1 : S$ **do**
      Sample $\mathbf{y}^{s,k} \sim p(\mathbf{y}|\boldsymbol{\theta}_k)$ ;
    $\hat{E}(\phi(\mathbf{y})) = \frac{1}{S}\sum_{s=1}^{S} \phi(\mathbf{y}^{s,k})$;
    **for** *each training case* $i$ *in minibatch* **do**
      $\mathbf{g}_{ik} = \phi(\mathbf{y}_i) - \hat{E}(\phi(\mathbf{y}))$ ;
    $\mathbf{g}_k = \frac{1}{B}\sum_{i \in B} \mathbf{g}_{ik}$;
    $\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \eta \mathbf{g}_k$;
    $k = k + 1$;
    Decrease step size $\eta$;

Source: Kevin Murphy, "Machine Learning: a probabilistic perspective"

# Stochastic Maximum Likelihood

**Algorithm :** Stochastic maximum likelihood for fitting an MRF

Initialize weights $\boldsymbol{\theta}$ randomly;
$k = 0$, $\eta = 1$ ;
**for** *each epoch* **do**
  **for** *each minibatch of size* $B$ **do**
    **for** *each sample* $s = 1 : S$ **do**
      Sample $\mathbf{y}^{s,k} \sim p(\mathbf{y}|\boldsymbol{\theta}_k)$ ;
    $\hat{E}(\phi(\mathbf{y})) = \frac{1}{S}\sum_{s=1}^{S} \phi(\mathbf{y}^{s,k})$;
    **for** *each training case* $i$ *in minibatch* **do**
      $\mathbf{g}_{ik} = \phi(\mathbf{y}_i) - \hat{E}(\phi(\mathbf{y}))$ ;
    $\mathbf{g}_k = \frac{1}{B}\sum_{i \in B} \mathbf{g}_{ik}$;
    $\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \eta \mathbf{g}_k$;
    $k = k + 1$;
    Decrease step size $\eta$;

// split the observed data $\mathrm{y}_i, \forall i = 1 \ldots N$ into sets of size $B$

// draw S samples from the model's distribution $p(y|\theta_k)$

// note that we fixed the parameter at $\theta_k$ (current estimate)

* We will discuss more about MCMC sampling in Lecture 11

Source: Kevin Murphy, "Machine Learning: a probabilistic perspective"

# Stochastic Maximum Likelihood

**Algorithm :** Stochastic maximum likelihood for fitting an MRF

Initialize weights $\boldsymbol{\theta}$ randomly;
$k = 0, \eta = 1$ ;
**for** *each epoch* **do**
  **for** *each minibatch of size* $B$ **do**
    **for** *each sample* $s = 1 : S$ **do**
      Sample $\mathbf{y}^{s,k} \sim p(\mathbf{y}|\boldsymbol{\theta}_k)$ ;
    $\hat{E}(\phi(\mathbf{y})) = \frac{1}{S}\sum_{s=1}^{S} \phi(\mathbf{y}^{s,k})$;
    **for** *each training case* $i$ *in minibatch* **do**
      $\mathbf{g}_{ik} = \phi(\mathbf{y}_i) - \hat{E}(\phi(\mathbf{y}))$ ;
    $\mathbf{g}_k = \frac{1}{B}\sum_{i \in B} \mathbf{g}_{ik}$;
    $\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \eta \mathbf{g}_k$;
    $k = k + 1$;
    Decrease step size $\eta$;

// split the observed data $\mathrm{y}_i, \forall i = 1 \ldots N$ into sets of size $B$

// draw S samples from the posterior distribution $p(y|\theta_k)$

// note that we fixed the parameter at $\theta_k$ (current estimate)

// compute the approximated gradient, $g_k \approx \frac{\partial l}{\partial \theta}$

Source: Kevin Murphy, "Machine Learning: a probabilistic perspective"

# Stochastic Maximum Likelihood

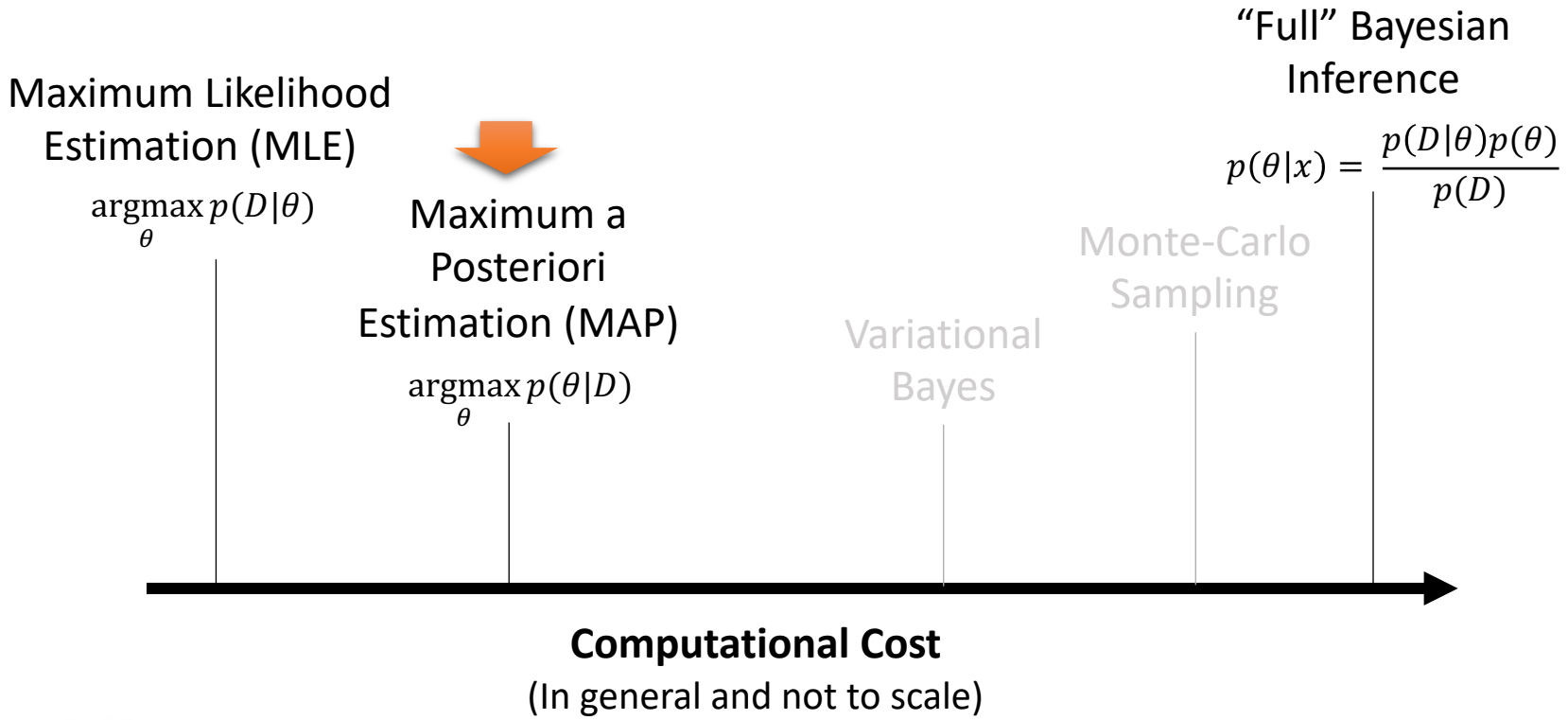
**Algorithm :** Stochastic maximum likelihood for fitting an MRF

Initialize weights $\boldsymbol{\theta}$ randomly;
$k = 0$, $\eta = 1$ ;
**for** *each epoch* **do**
  **for** *each minibatch of size* $B$ **do**  // split the observed data $y_i, \forall i = 1 \ldots N$ into sets of size $B$
    **for** *each sample* $s = 1 : S$ **do**
      Sample $\mathbf{y}^{s,k} \sim p(\mathbf{y}|\boldsymbol{\theta}_k)$ ;  // draw S samples from the posterior distribution $p(y|\theta_k)$
    $\hat{E}(\phi(\mathbf{y})) = \frac{1}{S}\sum_{s=1}^{S} \phi(\mathbf{y}^{s,k})$;  // note that we fixed the parameter at $\theta_k$ (current estimate)
    **for** *each training case* $i$ *in minibatch* **do**
      $\mathbf{g}_{ik} = \phi(\mathbf{y}_i) - \hat{E}(\phi(\mathbf{y}))$ ;  // compute the approximated gradient, $g_k \approx \frac{\partial l}{\partial \theta}$
    $\mathbf{g}_k = \frac{1}{B}\sum_{i \in B} \mathbf{g}_{ik}$;
    $\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \eta \mathbf{g}_k$;  // stochastic gradient descent update step
    $k = k + 1$;
    Decrease step size $\eta$;

Source: Kevin Murphy, "Machine Learning: a probabilistic perspective"

# Learning Parameters

- Common approaches to learn the unknown parameters $\theta$ from a set of given data $\mathcal{D} = \{x[1], \ldots, x[N]\}$:



# Maximum A Posteriori (MAP): MRF (UGM)

- We can also do MAP to learn the unknown parameters in UGM, where we add a prior term:

$$\underset{\theta}{\operatorname{argmax}} \left\{ \sum_i \log p(y_i|\theta) + \underbrace{\log p(\theta)}_{\text{Prior term}} \right\}$$

- A Gaussian prior is often used:

$$p(\theta) = \mathcal{N}(\theta|\mu, \Sigma) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp[-\tfrac{1}{2}(\theta-\mu)^T\Sigma^{-1}(\theta-\mu)]$$

- Where $(\mu, \Sigma)$ are the hyperparameters.


# Recap

# Learning Outcomes

- Students should be able to:

1. Explain the concepts of Markov properties (global, local and pairwise) and use it to find all conditional independences in an UGM.

2. Use clique potential functions to parameterize a Markov Random Field, i.e. to represent the joint distribution with clique potential functions.

3. Learn the parameters in a MRF (stochastic) gradient-based methods.

# Undirected Graphical Models:
## In a nutshell

- An alternative to DGMs is to use an Undirected Graphical model (UGM), also called a Markov Random Field (MRF) or Markov network.

- Formally, an UGM is a graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$, where:
  - $\mathcal{V}$ is a set of nodes that are in one-to-one correspondence with a set of random variables.
  - $\mathcal{E}$ is a set of undirected edges.

# Undirected Graphical Models:
## In a nutshell

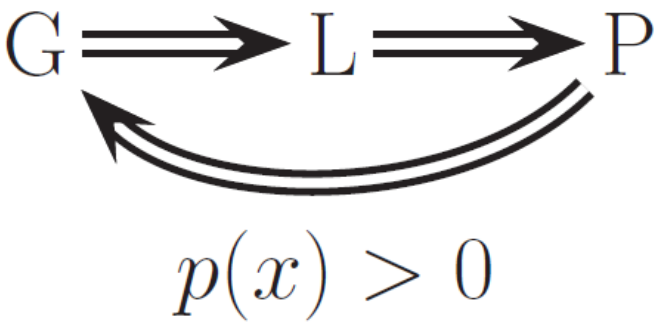
- Parameterization is achieved via factors.
- A factor $\varphi(\mathcal{C})$ is a function that maps a set of random variables $\mathcal{C} = \{X, \ldots, Z\}$ to a real number.
  - Restrict: non-negative factors only ($\varphi(\mathcal{C})$ only maps to non-negative numbers)
- The factorization is:

$$p(x_1, \ldots, x_N) = \frac{1}{Z} \prod_{j=1}^{M} \varphi_j(\mathcal{C}_j)$$

Compare to DGM:

$$p(x_1, \ldots, x_N) = \prod_{i=1}^{N} p(x_i | x_{\pi_i})$$

# Conditional Independence




- Easy to check that global Markov implies local Markov which implies pairwise Markov.
- What is less obvious, but true (assuming positive distributions $p(\mathbf{x}) > 0$ for all $\mathbf{x}$), is that pairwise Markov implies global Markov.

Image Source: "Machine Learning – A Probabilistic Perspective", Kevin Murphy

# Questions about MRFs

- Is the MRF correct/sound?
  - Does a conditional independence identified by separation always exist in the distribution? **Yes.** ☺
- Is the MRF complete?
  - If a conditional independence exists in the distribution, can it always be detected by separation? **Almost Yes.** 😐
- How expressive are MRF as a modeling language?
  - Can they exactly represent all conditional independencies for a given distribution?

**No.** ☹

# Parameter Learning: UGM (MRF)

- Consider a Markov Random Field (MRF) in log-linear form, where $c$ indexes the cliques:

$$p(\mathbf{y}|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \exp\left(\sum_c \boldsymbol{\theta}_c^T \boldsymbol{\phi}_c(\mathbf{y})\right)$$

- The scaled log-likelihood is given by:

$$\ell(\boldsymbol{\theta}) \triangleq \frac{1}{N} \sum_{i}^{N} \log p(\mathbf{y}_i|\boldsymbol{\theta}) = \frac{1}{N} \sum_{i}^{N} \left[\sum_c \boldsymbol{\theta}_c^T \boldsymbol{\phi}_c(\mathbf{y}_i) - \log Z(\boldsymbol{\theta})\right]$$

# Conditional Random Fields

- A CRF or discriminative random field, is just a version of an MRF where all the clique potentials are conditioned on input features X:

$$p(\mathbf{y}|\mathbf{x},\mathbf{w}) = \frac{1}{Z(\mathbf{x},\mathbf{w})}\prod_c \psi_c(\mathbf{y}_c|\mathbf{x},\mathbf{w})$$

- We will usually assume a log-linear representation of the potentials:

$$\psi_c(\mathbf{y}_c|\mathbf{x},\mathbf{w}) = \exp(\mathbf{w}_c^T \boldsymbol{\phi}(\mathbf{x},\mathbf{y}_c))$$

- where $\phi(\mathrm{x}, y_c)$ is a feature vector derived from the global inputs $\mathrm{X}$ and the local set of labels $Y_c$.

# Parameter Learning: UGM (CRF)

- Consider a Conditional Random Field (CRF) in log-linear form, where $c$ indexes the cliques:

$$p(\mathbf{y}|\mathbf{x},\mathbf{w}) = \frac{1}{Z(\mathbf{x},\mathbf{w})}\prod_{c}\exp(\mathbf{w}_c^T\phi_c(\mathbf{x},\mathbf{y}_c))$$

- $\phi_c(\mathrm{x}, y_c)$ is a feature vector derived from the global inputs $\mathrm{x}$ and the local set of labels $y_c$.

# Learning Outcomes

- Students should be able to:

1. Explain the concepts of Markov properties (global, local and pairwise) and use it to find all conditional independences in an UGM.

2. Use clique potential functions to parameterize a Markov Random Field, i.e. to represent the joint distribution with clique potential functions.

3. Learn the parameters in a MRF (stochastic) gradient-based methods.