NUS National University of Singapore | School of Computing

CS5340
Uncertainty Modeling in AI

Lecture 4:
Markov Random Fields
(Undirected Graphical Models)

Asst. Prof. Lee Gim Hee

AY 2018/19

Semester 1

Course Schedule

Week	Date	Topic	Remarks
1	15 Aug	Introduction to probabilities and probability distributions	
2	22 Aug	Fitting probability models	Hari Raya Haji*
3	29 Aug	Bayesian networks (Directed graphical models)	
4	05 Sep	Markov random Fields (Undirected graphical models)	
5	12 Sep	I will be traveling	No Lecture
6	19 Sep	Variable elimination and belief propagation	
-	26 Sep	Recess week	No lecture
7	03 Oct	Factor graph and the junction tree algorithm	
8	10 Oct	Parameter learning with complete data	
9	17 Oct	Mixture models and the EM algorithm	
10	24 Oct	Hidden Markov Models (HMM)	
11	31 Oct	Monte Carlo inference (Sampling)	
12	07 Nov	Variational inference	
13	14 Nov	Graph-cut and alpha expansion	

* **Make-up lecture**: 25 Aug (Sat), 9.30am-12.30pm, LT 15

Acknowledgements

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

1. "Machine learning - a probabilistic approach", Kevin Murphy (Chapter 19)

2. "Probabilistic graphical models", Koller and Friedman (Chapter 4)

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

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. Describe the differences and similarities between a Markov Random Field and Conditional Random Field.

Why Undirected Graphical Models?

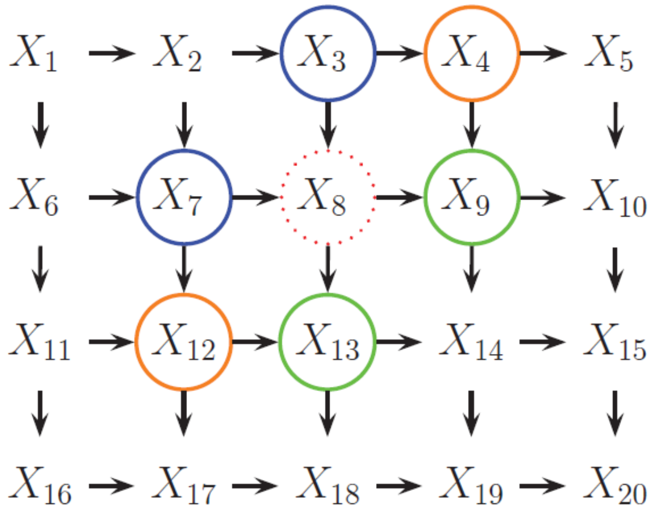
- We discussed the Directed Graphical Models (DGMs) or Bayesian Networks in the last lecture.

- However, for some domains, the requirement for a directed edge is rather awkward.

Why Undirected Graphical Models?

Example:

Causal MRF



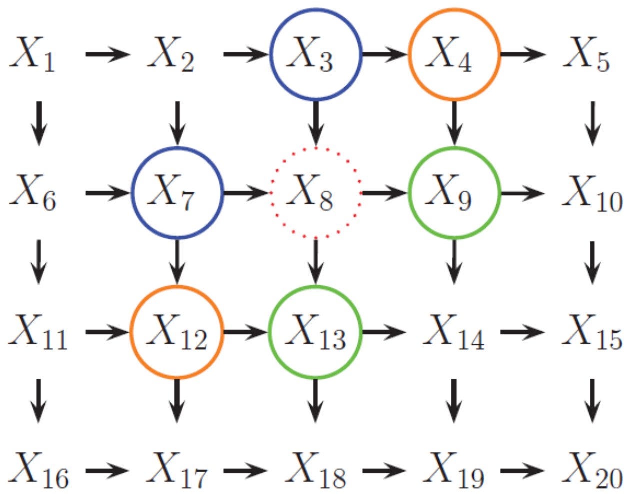
- Modeling a 2D image where the intensity of neighboring pixels are correlated.
- We can create a DAG model with a 2d lattice topology known as a causal MRF or a Markov mesh.

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

Why Undirected Graphical Models?

Example:

Causal MRF



- However, its conditional independence properties are rather unnatural.

- The Markov blanket of the node X_8 in the middle is the other colored nodes (3, 4, 7, 9, 12 and 13) rather than just its 4 nearest neighbors as one might expect.

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

Why Undirected Graphical Models?

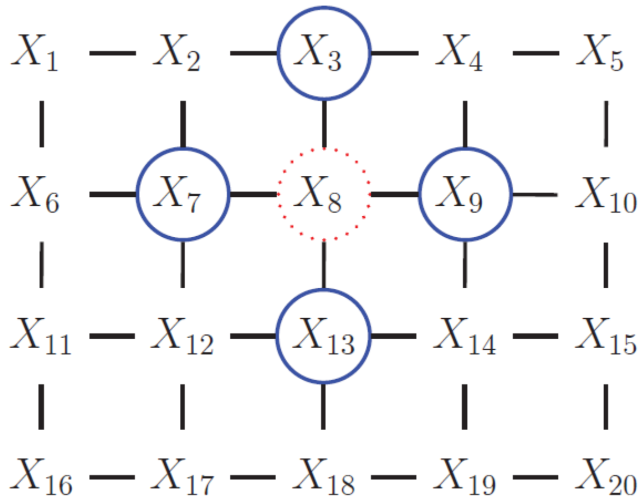
- An alternative is to use an Undirected Graphical model (UGM), also called a Markov Random Field (MRF) or Markov network.

- 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.

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

Why Undirected Graphical Models?

Example:



- We use an undirected 2d lattice to model a 2D image where the intensity of neighboring pixels are correlated.
- Now the Markov blanket of each node is just its nearest neighbors (more on Markov blanket for UGMs later).

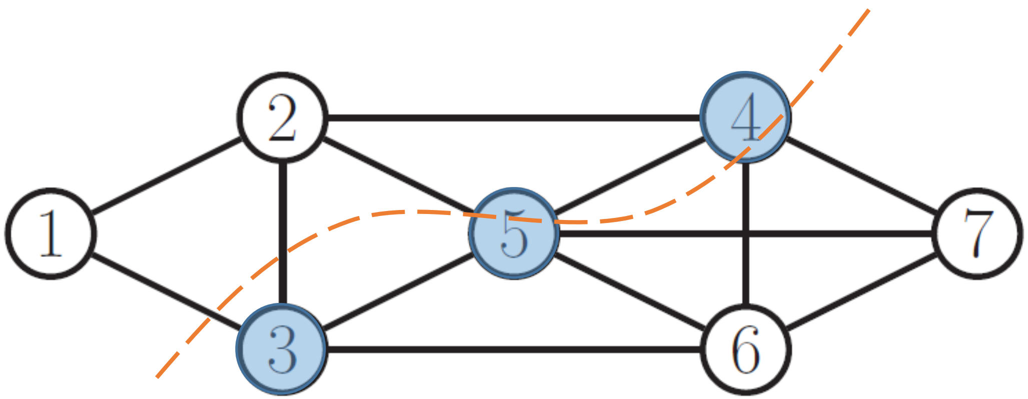
Image Source: "Machine Learning – A Probabilistic Perspective", Kevin Murphy

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 paths 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

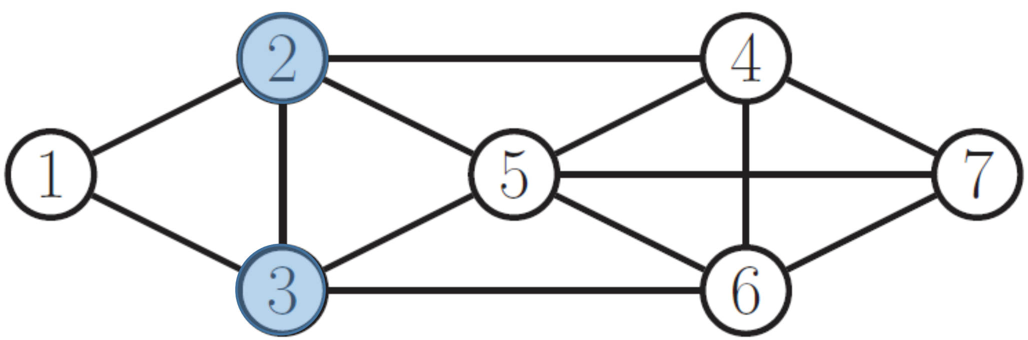
2. Local Markov Property

- The set of nodes that renders a node X_s conditionally independent of all the other nodes in $\mathcal{G}$:

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

- This is called X_s' Markov blanket denote by $\mathrm{mb}(X_s)$.

Example:



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

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

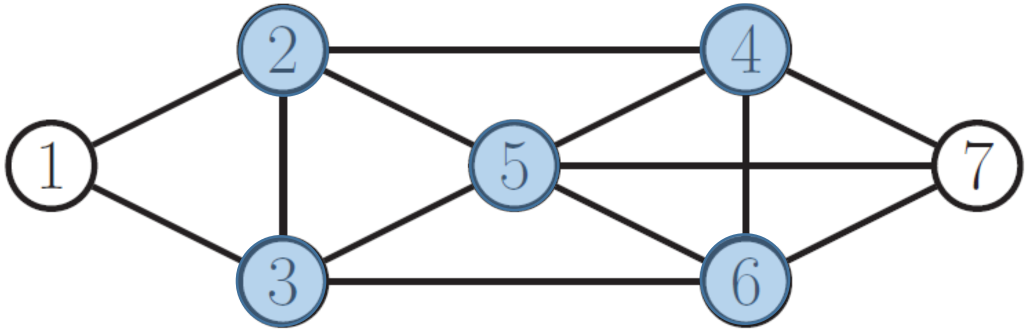
Conditional Independence

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

Conditional Independence

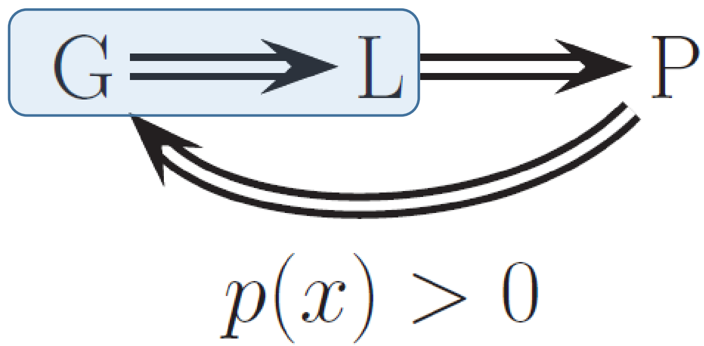
$$
\begin{tikzcd}
G \arrow[r, Rightarrow] & L \arrow[r, Rightarrow] & P \arrow[ll, Rightarrow, bend left, "p(x) > 0"]
\end{tikzcd}
$$

- Obvious that global Markov implies local Markov which implies pairwise Markov.

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

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

Conditional Independence



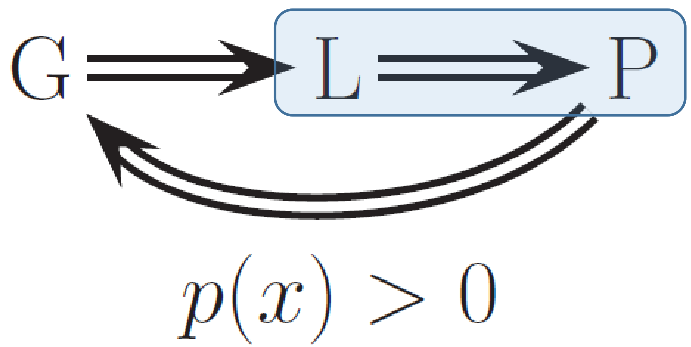
Proof:

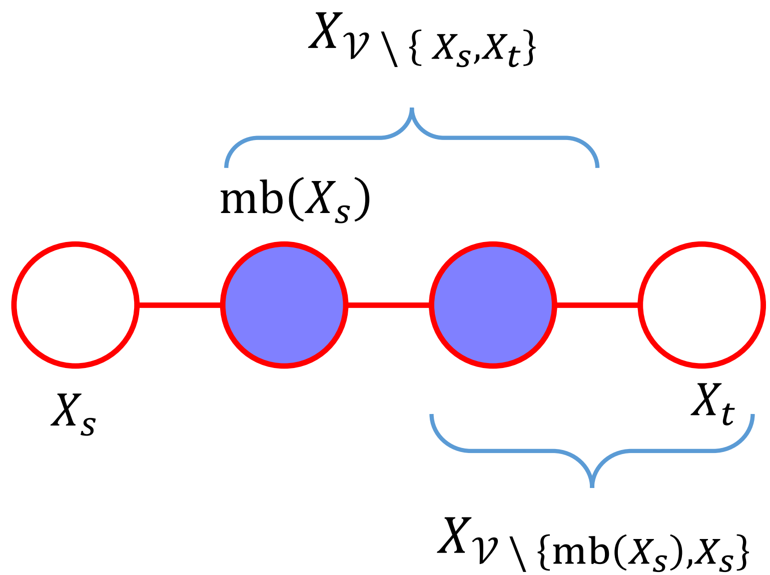
The global Markov property implies the local Markov property: this is the case when the sets $X_A = X_s,\ X_C = \mathrm{mb}(X_s)$, and $X_B = \{\mathcal{V} \setminus \{\mathrm{mb}(X_s), X_s\}$.

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

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

Conditional Independence





Proof:

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

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

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

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

Intersection Lemma

For positive distributions, and for mutually disjoint sets X, Y, W, Z :

If $X \perp Y \mid \{W, Z\}$ and $X \perp W \mid \{Y, Z\} \Rightarrow X \perp \{Y, W\} \mid Z$

Proof:

From $X \perp Y \mid \{Z, W\}$ and $X \perp W \mid \{Z, Y\}$, we can write the joint distribution $p(X, Y, Z, W)$ as:

$$\overbrace{f_{XWZ}(X, W, Z) f_{WYZ}(W, Y, Z)}^{X \perp Y \mid \{Z, W\}} = \overbrace{g_{XYZ}(X, Y, Z) g_{W,Y,Z}(W, Y, Z)}^{X \perp W \mid \{Z, Y\}}$$

$$\Big\Downarrow \text{ positive distributions}$$

$$p(X, Y, Z, W) = \mu_{XZ}(X, Z) \mu_{W,Y,Z}(W, Y, Z) \Rightarrow X \perp \{Y, W\} \mid Z$$

Non-Unique Probability Factorization

- Both

$$f_{XWZ}(X, W, Z) f_{WYZ}(W, Y, Z) \qquad \text{and}$$
$$g_{XYZ}(X, Y, Z) g_{W,Y,Z}(W, Y, Z)$$

are valid factorizations of the joint distribution

$$p(X, Y, Z, W).$$

- Due to non-uniqueness of probability factorization, which can be explained by the "Independence-Map"!

Independence-Map

- Also known as the I-Map.
- The I-Map of a joint distribution $p(x_1, \dots, x_N)$, often written as $I(p)$ represents all independencies in $p(x_1, \dots, x_N)$.
- Similarly, the I-Map of a directed/undirected graph $\mathcal{G}$, i.e. $I(\mathcal{G})$ represents all independencies encoded in $\mathcal{G}$.
- $\mathcal{G}$ is a valid representation of p if $I(\mathcal{G}) \subseteq I(p)$.

Independence-Map

- **Given:** 4 disjoint sets W, X, Y, Z, where the non-zero distribution $p(X, Y, Z, W)$ contains **at least two** conditional independences.

- I-map implies that all the following are valid factorizations of the joint distribution $p(X, Y, Z, W)$:

$$p(X,Y,Z,W) = f_{XWZ}(X,W,Z)f_{WYZ}(W,Y,Z),$$

for the conditional independence $X \perp Y \mid \{Z, W\}$

$$p(X,Y,Z,W) = g_{XYZ}(X,Y,Z)g_{W,Y,Z}(W,Y,Z)$$

for the conditional independence $X \perp W \mid \{Z, Y\}$

Independence-Map

- This is because $p(X, Y, Z, W)$ contains at least two conditional independences, i.e. $X \perp Y \mid \{Z, W\}$ and $X \perp W \mid \{Z, Y\}$.

- And the respective factorizations encodes only one conditional independence each, i.e. $I(f) \subseteq I(p)$ and $I(g) \subseteq I(p)$.

Intersection Lemma

- We know that $p(X, Y, Z, W)$ has at least two conditional independences, hence, there exists another factorization that satisfies BOTH the conditional independences.

- Since

$$p(X, Y, Z, W) = f_{XWZ}(X, W, Z) f_{WYZ}(W, Y, Z),$$
$$p(X, Y, Z, W) = g_{XYZ}(X, Y, Z) g_{W,Y,Z}(W, Y, Z)$$

represent the same distribution $p(X, Y, Z, W)$, we can equate them, i.e.

$$f_{XWZ}(X, W, Z) f_{WYZ}(W, Y, Z) = g_{XYZ}\ (X, Y, Z) g_{W,Y,Z}(W, Y, Z)$$

Intersection Lemma

$$f_{XWZ}(X,W,Z)f_{WYZ}(W,Y,Z) = g_{XYZ}\ (X,Y,Z)g_{W,Y,Z}(W,Y,Z)$$

Through inspection, we see that $\{X, Z\}$ and $\{W, Y, Z\}$ have to appear in two factors, i.e.

$$p(X,Y,Z,W) = \mu_{XZ}(X,Z)\mu_{W,Y,Z}(W,Y,Z)$$

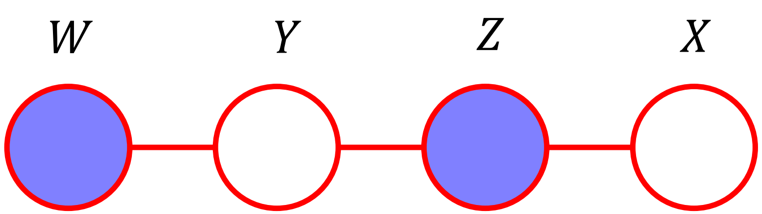
As a result, we get both the conditional independences $X \perp Y \mid \{Z, W\}$ and $X \perp W \mid \{Z, Y\}$ encoded in the same factorization.

In addition, we observe an additional conditional independence $X \perp \{Y, W\} \mid Z$, which is the intersection lemma.

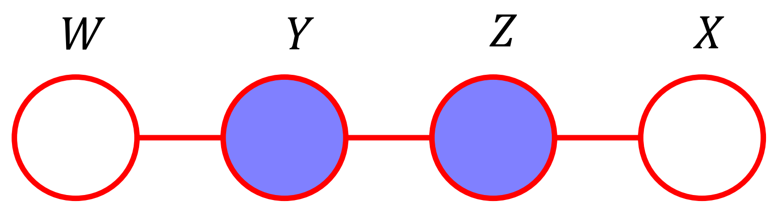
Intersection Lemma

For positive distributions, and for mutually disjoint sets X, Y, W, Z :

If $X \perp Y \mid \{W, Z\}$ and $X \perp W \mid \{Y, Z\} \Rightarrow X \perp \{Y, W\} \mid Z$

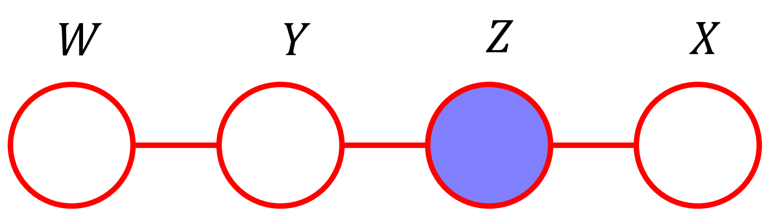


$X \perp Y \mid \{W, Z\}$



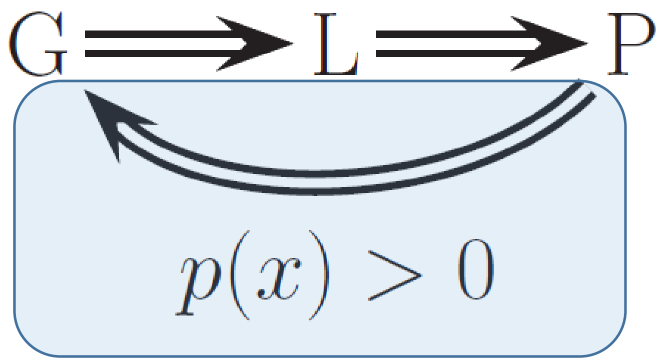
$X \perp W \mid \{Y, Z\}$

positive distributions



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

Conditional independence

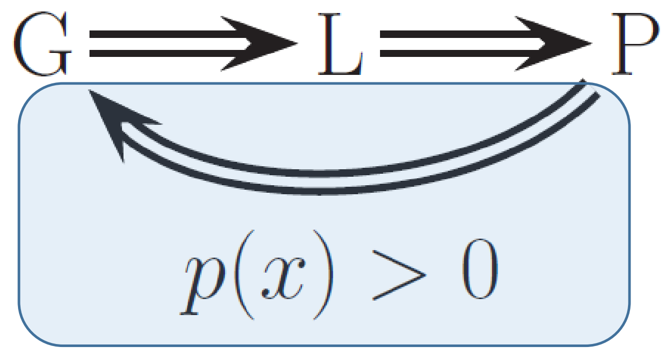


Proof:

Let $S, A, B, D \subset \mathcal{V}$ be disjoint sets of nodes with S separating A from B in the graph $\mathcal{G}$, where $A \notin \emptyset$ and $B \notin \emptyset$. We will prove that pairwise Markov implies global Markov using backward induction.

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

Conditional independence



Proof:

Let $d = |\mathcal{V}|$, when $|S| = d - 2$:

$A \perp B \mid S,$ where $|A| = |B| = 1$

$\Longrightarrow$ pairwise Markov

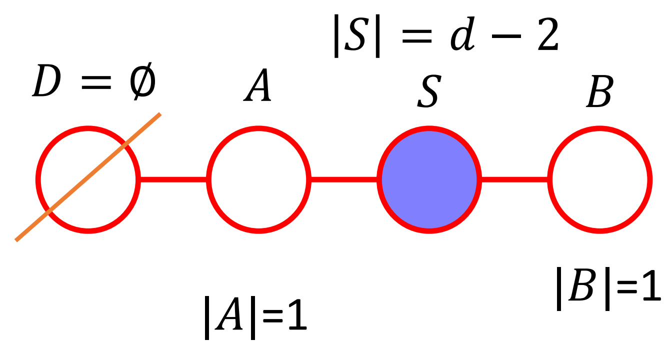
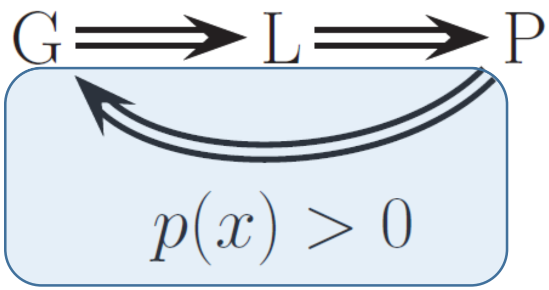


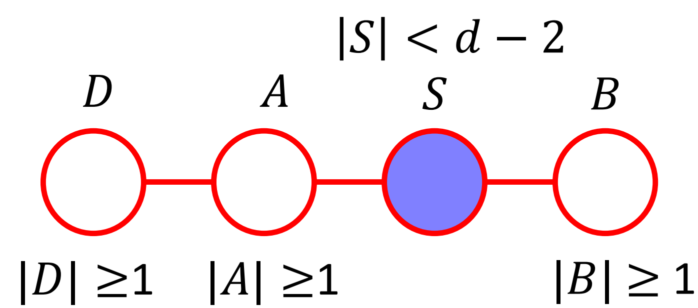
Image Source: “Machine Learning – A Probabilistic Perspective”, Kevin Murphy

Conditional independence



Proof:

For $|S| < d - 2$, WLOG, let us assume the set of nodes D is connected only to A, where $|D| \geq 1$, $|A| \geq 1$ and $|B| \geq 1$. We have:



$$A \perp B \mid \{S, D\} \text{ and } B \perp D \mid \{A, S\}$$

Intersection Lemma

$$B \perp \{A, D\} \mid S \quad \Longrightarrow \text{global Markov}$$

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

Conditional independence

$$\begin{tikzcd} G \arrow[r, Rightarrow] & L \arrow[r, Rightarrow] & P \arrow[ll, Rightarrow, bend left, "p(x) > 0"] \end{tikzcd}$$

- The importance of this result is that it is usually easier to empirically assess pairwise conditional independence.
- Such pairwise CI statements can be used to construct a graph from which global CI statements can be extracted.

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

Comparative Semantics

- We have seen that it is easier to determine conditional independence using UGMs than DGMs.

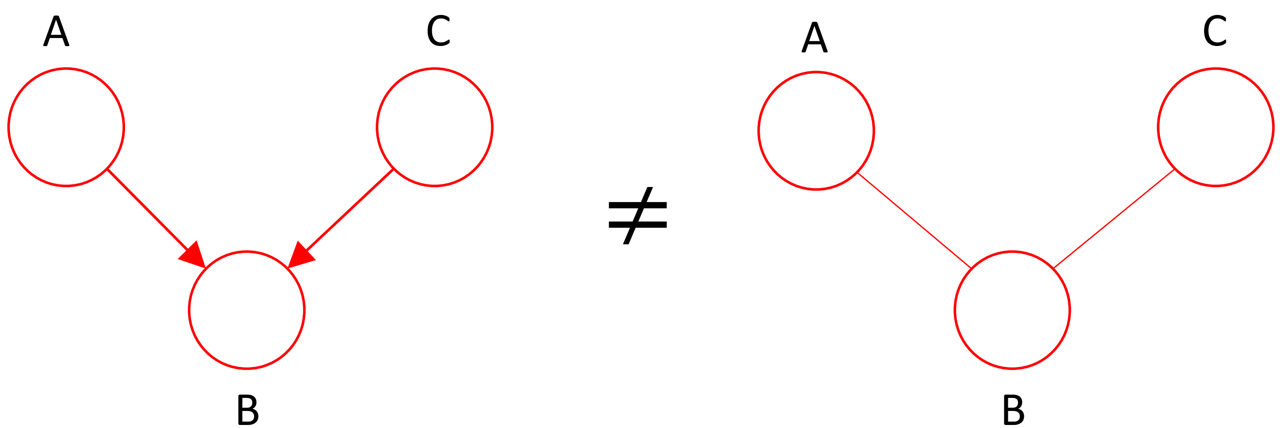
- **Question**: Can we determine conditional independence in a DGM using a UGM, or vice versa?

- This is NOT possible in general!

Comparative Semantics

- It is tempting to simply convert the DGM to a UGM by dropping the orientation of the edges, but this is not always correct!

Example:



$$A \perp C$$

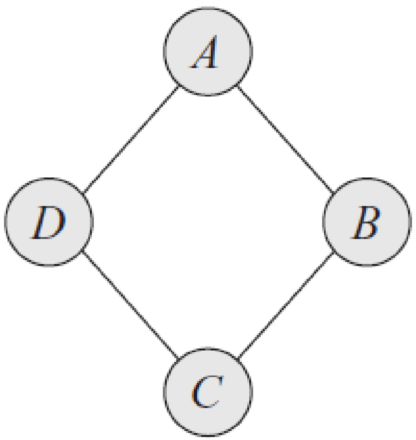
$$A \perp C \mid B$$

This conditional independence is NOT in the DGM!

Comparative Semantics

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

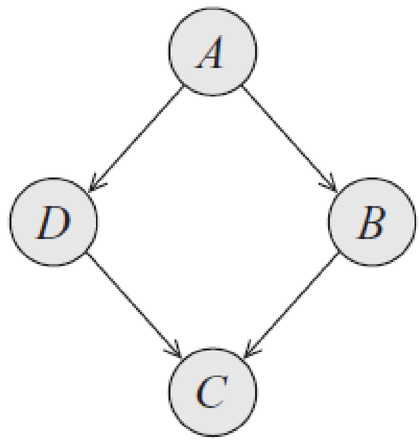
UGM



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

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

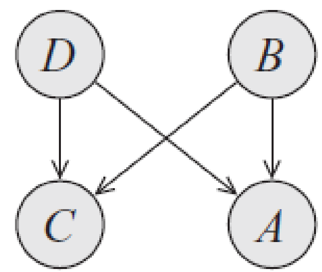
Attempt 1



✓ $A \perp C \mid \{B, D\}$

✗ $B \perp D \mid A$

Attempt 2



✓ $A \perp C \mid \{B, D\}$

✗ $B \perp D$

Image source: Koller and Friedman 2009

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.

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

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.

Parameterization of MRFs

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

- Retain the ability the all-important representation of the joint as a product of local functions.

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 cannot have a local function that depends on both X_i and X_j.

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

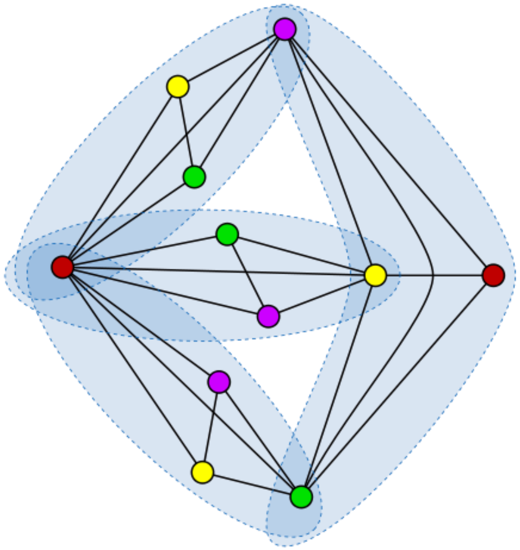
Parameterization of MRFs

How do we decide the domain of the local functions?

- Our argument thus far suggested 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.



Conditional independence is impossible for any two nodes in a maximal clique!

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

Parameterization of MRFs

Hammersley-Clifford theorem:

A positive distribution $p(y) > 0$ satisfies the CI properties of an undirected graph $\mathcal{G}$ 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
- θ is the parameter of the factors $\psi_c(.)$ for $c \in \mathcal{C}$
- $Z(\theta)$ is the partition function

Parameterization of MRFs

Hammersley-Clifford theorem:

$Z(\theta)$ is the partition function given by:

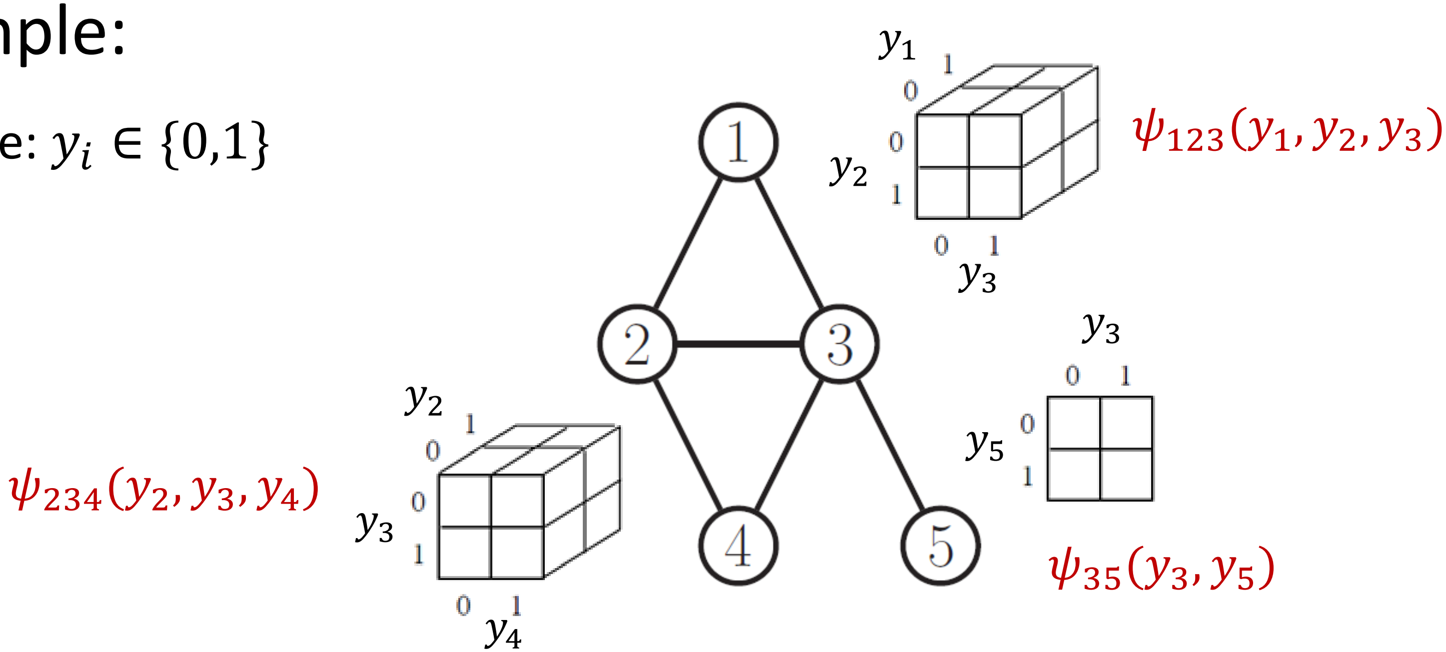
$$Z(\boldsymbol{\theta}) \triangleq \sum_{y} \prod_{c \in \mathcal{C}} \psi_c(\mathbf{y}_c | \boldsymbol{\theta}_c)$$

Since $\psi(.)$ can be any arbitrary positive function, the partition function $Z(\theta)$ ensures the overall distribution sums to 1.

Parameterization of MRFs

Example:

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



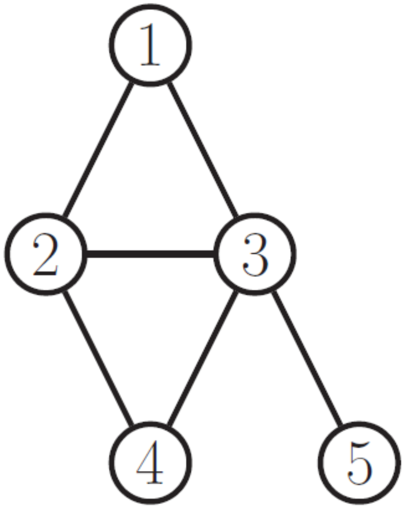
$$p(\mathbf{y}|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \psi_{123}(y_1, y_2, y_3)\psi_{234}(y_2, y_3, y_4)\psi_{35}(y_3, y_5)$$

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

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.



Example 1:

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

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

Gibbs Distribution

- There is a deep connection between UGMs and statistical physics.
- In particular, there is a model known as the Gibbs distribution, which can be written as follows:

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

- $E(y_c) > 0$ is the energy associated with the variables in clique *c*.

Gibbs Distribution

- We can convert the Gibbs distribution to a UGM by defining:

$$\psi_c(\mathbf{y}_c|\boldsymbol{\theta}_c) = \exp(-E(\mathbf{y}_c|\boldsymbol{\theta}_c))$$

- We see that high probability states correspond to low energy configurations.

- Also known as energy based models, hence the term "potential function" for $\psi_c(.)$.

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:

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

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

Representing Potential Functions

- The resulting log probability has the form:

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

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

Representing Potential Functions

Example:

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

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

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:

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

$K^2 \times 1$

Representing Potential Functions

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.

Representing Potential Functions

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:

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

- *k* indexes the different features, corresponding to "ing", "qu-", etc., and ϕ_k is the corresponding binary feature function.

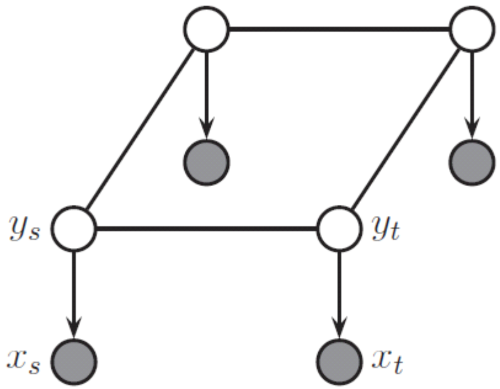
Representing Potential Functions

Example:

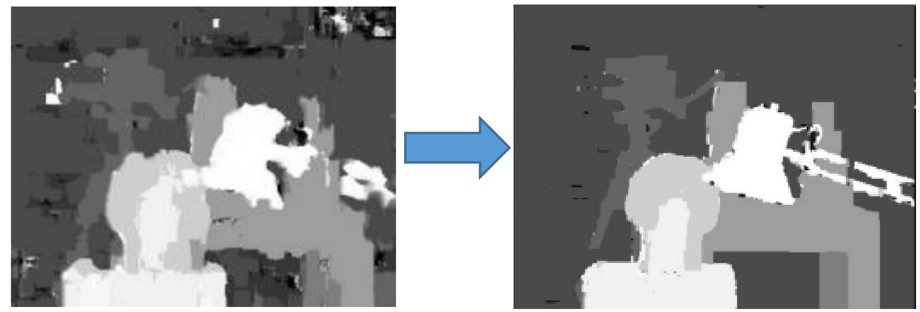
- By tying the parameters across locations, we can define the probability of a word of any length t using:

$$p(\mathbf{y}|\boldsymbol{\theta}) \propto \exp(\sum_t \sum_k \theta_k \phi_k(y_{t-1}, y_t, y_{t+1}))$$

Ising and Potts Models



Depth Map from Stereo Images



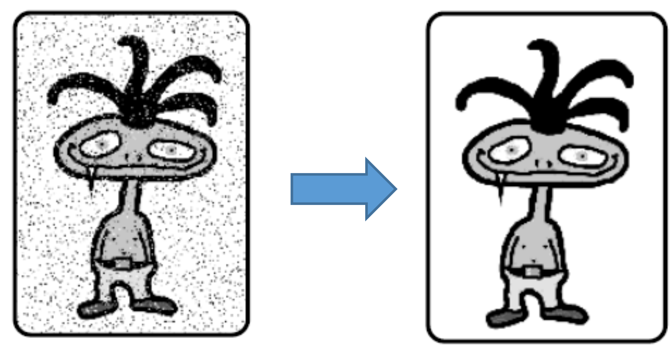
Observed Variables
$x \in \{1, \dots, L\}$

Latent Variables
$y \in \{1 \dots L\}$

$$p(y, x \mid J, \theta) = p(\mathbf{y}|J) \prod_t p(x_t|y_t, \boldsymbol{\theta})$$

$$= \underbrace{\left[\frac{1}{Z(J)} \prod_{s \sim t} \psi(y_s, y_t; J) \right]}_{\text{Pairwise potential}} \underbrace{\prod_t p(x_t|y_t, \boldsymbol{\theta})}_{\text{Unary potential}}$$

Binary Image Denoising

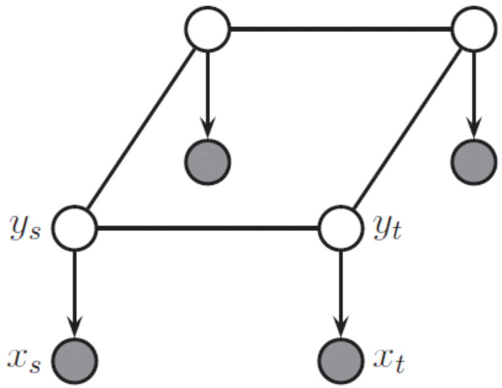


Observed Variables
$x \in \{0,1\}$

Latent Variables
$y \in \{0,1\}$

Image Source: “Computer Vision: Models, Learning, and Inference”, Simon Prince

Ising and Potts Models



$$p(y,x|J,\theta) = p(\mathbf{y}|J)\prod_t p(x_t|y_t,\boldsymbol{\theta})$$

$$= \underbrace{\left[\frac{1}{Z(J)}\prod_{s\sim t}\psi(y_s,y_t;J)\right]}_{\text{Pairwise potential}}\underbrace{\prod_t p(x_t|y_t,\boldsymbol{\theta})}_{\text{Unary potential}}$$

Ising Model:

$y_i \in \{0,1\}, \;\; x_i \in \{0,1\}$

$$E(y_s,y_t;J) = J|y_s - y_t|,$$
$$J > 0$$

Potts Model:

$y_i \in \{1,\ldots,L\}, \; x_i \in \{1,\ldots,L\}$

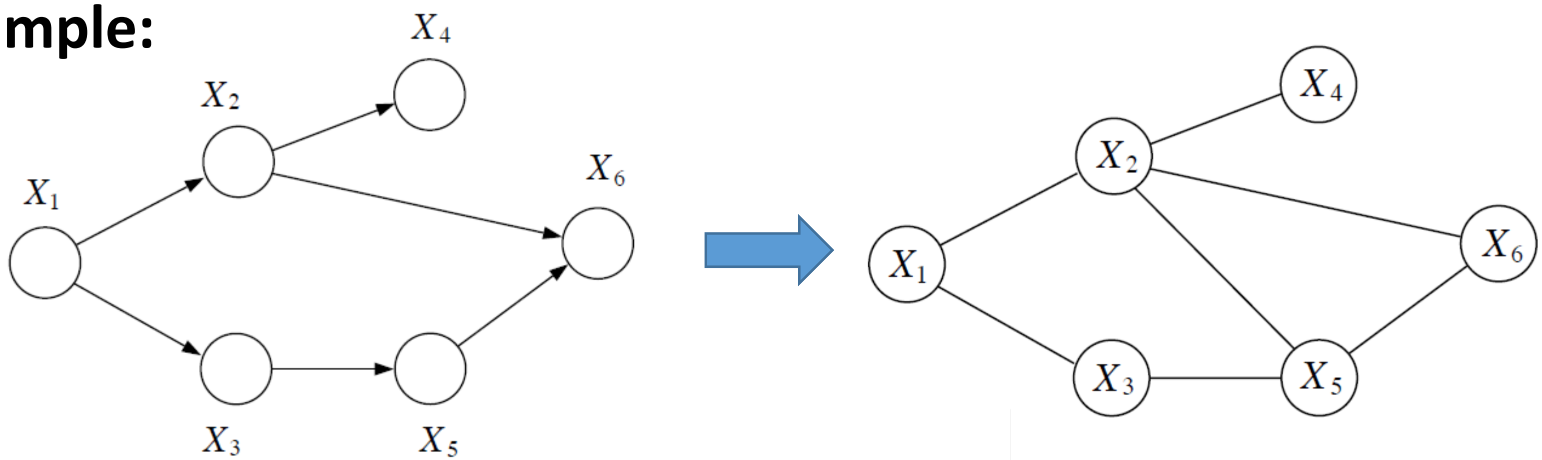
$$E(y_s,y_t;J) = J\min\left(|y_s - y_t|,1\right),$$
$$J > 0$$

Image Source: "Computer Vision: Models, Learning, and Inference", Simon Prince

Representing Potential Functions

- We can also use local conditional probabilities from a DGM to represent the potential functions in a UGM.

Example:



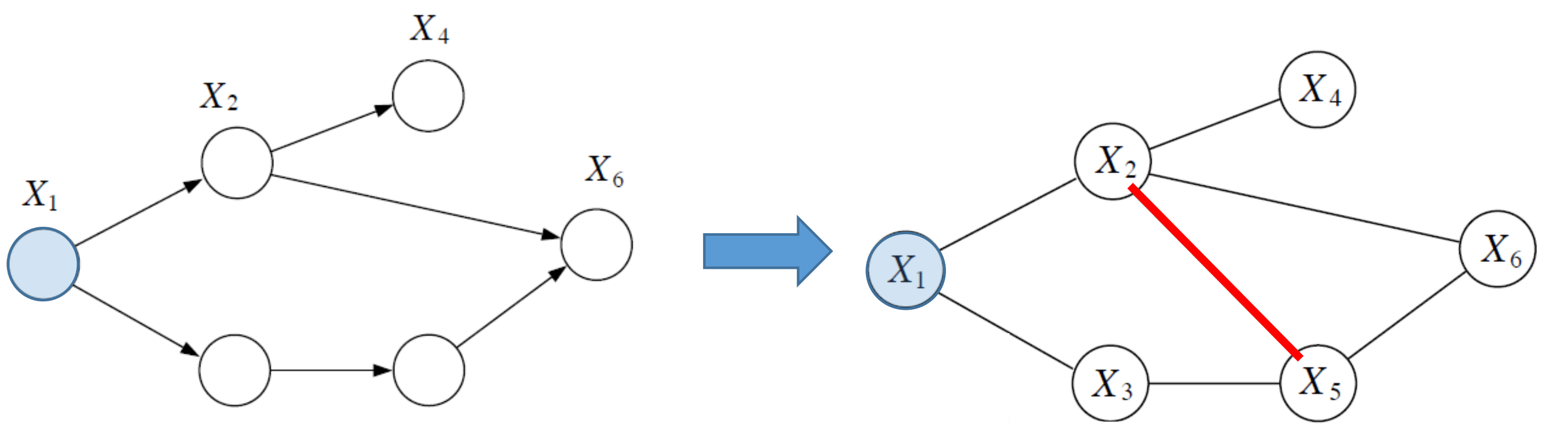
$$p(x) = \frac{1}{Z} \underbrace{\varphi_{12}(x_1, x_2)}_{p(x_2|x_1)} \underbrace{\varphi_{13}(x_1, x_3)}_{p(x_3|x_1)} \underbrace{\varphi_{14}(x_1, x_4)}_{p(x_4|x_1)} \underbrace{\varphi_{35}(x_3, x_5)}_{p(x_5|x_3)} \underbrace{\varphi_{256}(x_2, x_5, x_6)}_{p(x_6|x_2, x_5)}$$

$$Z = \sum_x \varphi_{12}(x_1, x_2)\varphi_{13}(x_1, x_3)\varphi_{14}(x_1, x_4)\varphi_{35}(x_3, x_5)\varphi_{256}(x_2, x_5, x_6) = \frac{1}{p(x_1)}$$

Image source: “An introduction to probabilistic graphical models”, Michael I. Jordan, 2002.

Moralization

- A DGM can be converted into a UGM by “marrying” the unmarried parents of a node, i.e. moralization.
- This process preserves the joint distribution, but conditional independence is lost!
- Moralization is important for exact inference (next lectures).



$x_2 \perp \{x_3, x_5,\} \mid x_1,$ $\qquad$ $x_2 \not\perp \{x_3, x_5,\} \mid x_1$

Image source: “An introduction to probabilistic graphical models”, Michael I. Jordan, 2002.

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

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

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

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 X and the local set of labels Y_c.

CRF vs MRF

Advantages:

1. No need to "waste resources" modeling things that we always observe.

   Focus our attention on modeling what we care about, i.e. the distribution of labels given the data.

2. We can make the potentials (or factors) of the model be data-dependent.

   e.g. in natural language processing problems, we can make the latent labels depend on global properties of the sentence, such as which language it is written in.

CRF vs MRF

Disadvantage:

1. Require labeled training data.

2. Learning is slower (more detail in the coming lectures).

Conditional Random Fields

Example: models for sequential data

Hidden Markov model:

$$p(\mathbf{x},\mathbf{y}|\mathbf{w}) = \prod_{t=1}^{T} p(y_t|y_{t-1},\mathbf{w})\underbrace{p(\mathbf{x}_t|y_t,\mathbf{w})}_{\text{Likelihood, i.e. generative}}$$

Chain structure MRF:

$$p(\mathrm{x},\mathrm{y} \mid \mathrm{w}) = \prod_{t=1}^{T} \underbrace{\psi(y_t;\ x_t,\mathrm{w})}_{\text{Likelihood, i.e. generative}} \prod_{t=1}^{T-1} \psi(y_t, y_{t+1}; \mathrm{w})$$

Chain structure CRF:

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

Posteriors, i.e. discriminative

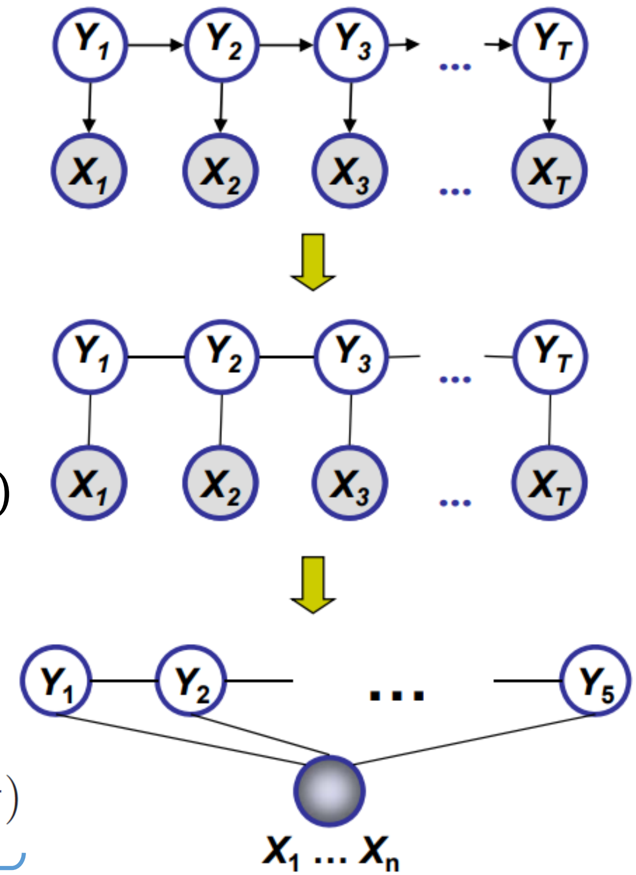
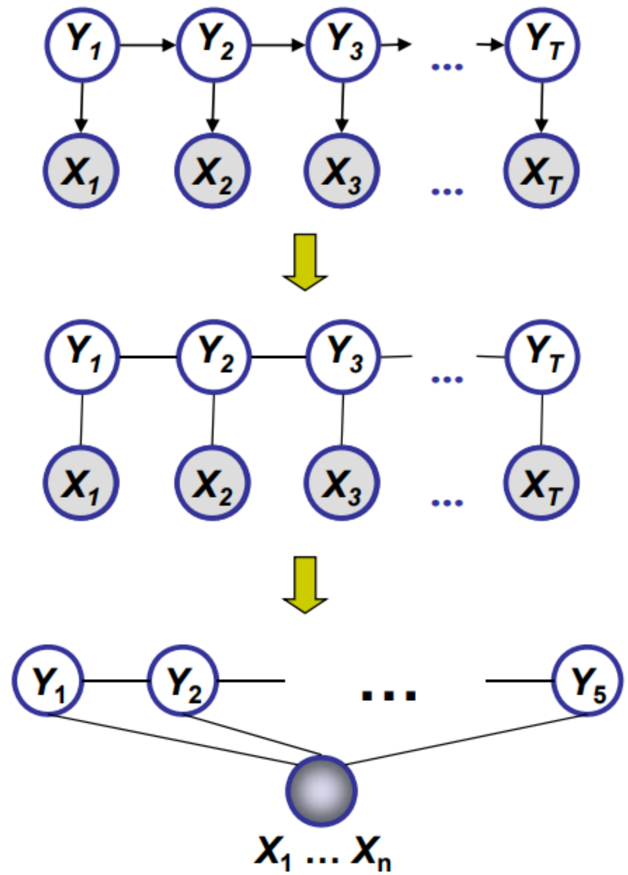


Image source: http://www.cs.cmu.edu/~epxing/Class/10708/lectures/lecture3-MRFrepresentation.pdf, Eric Xing

Conditional Random Fields

Example: models for sequential data

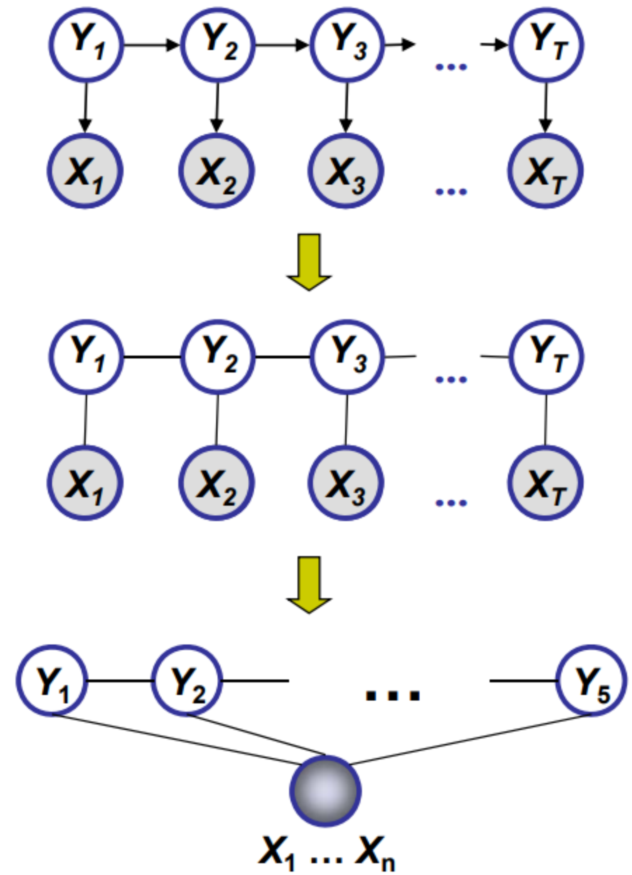


- HMM and MRF suffer from the label bias problem.
- Local features at time t do not influence states prior to time t.
- X_t is d-separated from all other nodes at Y_t thus blocking the information flow.

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

Conditional Random Fields

Example: models for sequential data

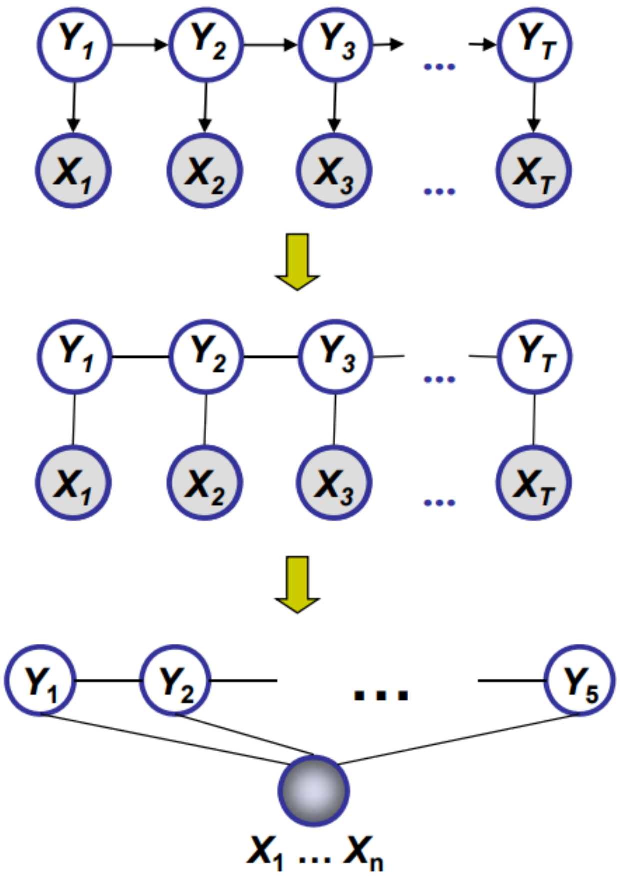


- Consider the part of speech (POS) tagging task.
- Suppose we see the word "banks".
- This could be a verb (as in "he banks at DBS"), or a noun (as in "the river banks were overflowing").
- Locally the POS tag for the word is ambiguous.

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

Conditional Random Fields

Example: models for sequential data



- Suppose that later in the sentence, we see the word "fishing".
- This gives us enough context to infer that the sense of "banks" is "river banks".
- However, in HMM and MRF the "fishing" evidence is d-separated.
- Problem is alleviated in CRF.

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