Undirected Graphical Models: Markov Random Fields

Probabilistic Graphical Models

Tavassolipour

Slides from Dr. Soleymani's PGM course, Sharif University of Technology

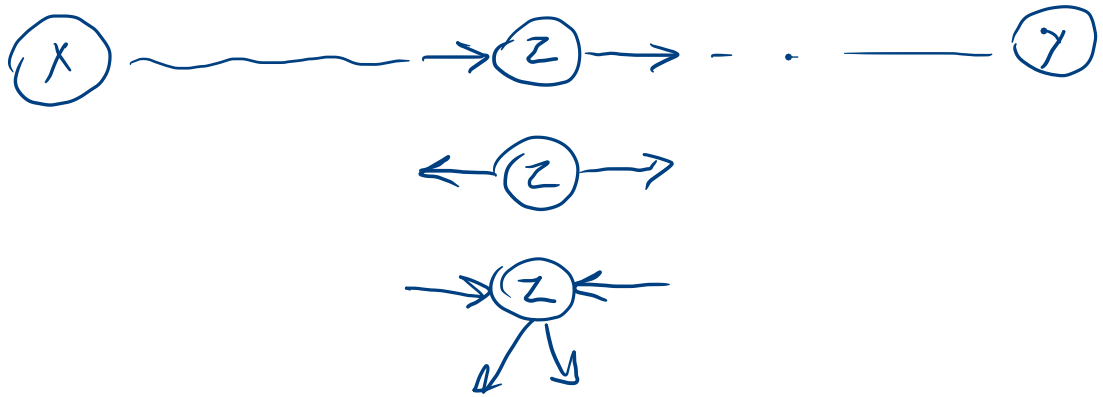
Bayesian Network:

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i \mid Pa(x_i))$$

$$\underline{I(G)} \subseteq I(p)$$

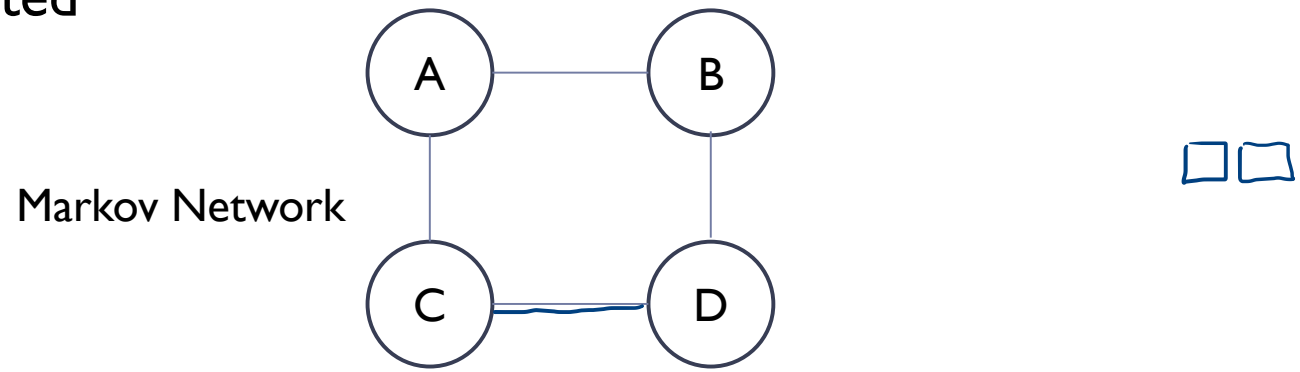
D-separation $\quad X \perp Y \mid Z$

↙ Directed



Markov Network

- Structure: ***undirected graph***
- Undirected edges show correlations (non-causal relationships) between variables
  - e.g., Spatial image analysis: intensity of neighboring pixels are correlated



MRF: Joint distribution

- Factor $\phi(X_1, \dots, X_k)$
  - $\phi: Val(X_1, \dots, X_k) \rightarrow \mathbb{R}$
  - Scope: $\{X_1, \dots, X_k\}$

- Gibbs Distribution

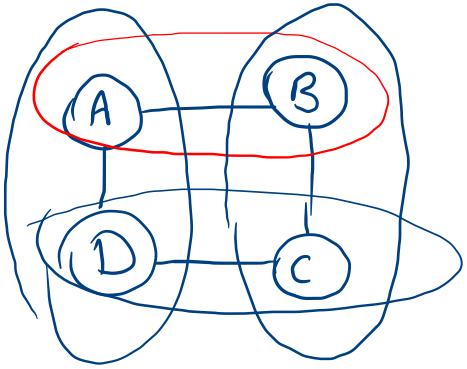
Joint distribution is parameterized by factors $\boldsymbol{\Phi} = \{\phi_1(\boldsymbol{D}_1), \dots, \phi_K(\boldsymbol{D}_K)\}$:

$$P(X_1, \dots, X_N) = \frac{1}{Z} \prod_k \phi_k(\boldsymbol{D}_k)$$

$\boldsymbol{D}_k$: the set of variables in the k-th factor

$$Z = \sum_{\boldsymbol{X}} \prod_k \phi_k(\boldsymbol{D}_k)$$

Z: normalization constant called **partition function**

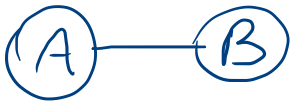


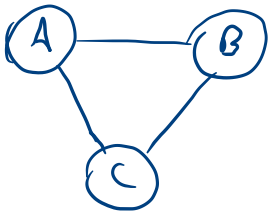
clique

$$P(A,B,C,D) = \frac{1}{Z}\,\phi_1(A,B)\,\phi_2(B,C)\,\phi_3(C,D)\,\phi_4(A,D)$$

normalization constant

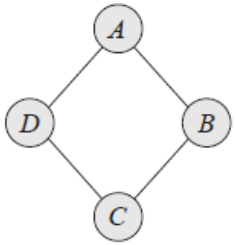
potential function





$P(A,B,C,D) = \frac{1}{Z} 30 \times 100 \times 1 \times 100$

Misconception example



[Koller & Friedman]

$\phi_1(A,B)$

$A = 0$

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

A	B	C	D	$\propto$
0	0	0	0	300000
0	0	0	1	
⋮				
1	1	1	1	

Factors show "compatibilities" between different values of the variables in their scope

A factor is only one contribution to the overall joint distribution.

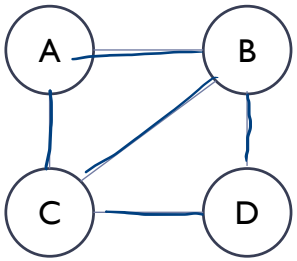
MRF Factorization: clique

Clique: subsets of nodes in the graph that are fully connected (complete subgraph)

Maximal clique: where no superset of the nodes in a clique are also compose a clique, the clique is maximal

Cliques:
{A,B,C}, {B,C,D}, {A,B}, {A,C}, {B,C}, {B,D}, {C,D}, {A}, {B}, {C}, {D}

Max-cliques:
{A,B,C}, {B,C,D}



$$P(A,B,C,D) = \frac{1}{Z}\Phi_1(A,B,C)\,\Phi_2(B,C,D)$$

$$P(A,B,C,D) = \frac{1}{Z}\Phi_1(A,B)\,\Phi_2(B,D)\,\Phi_3(C,D)\,\Phi_4(A,C)\,\Phi_5(B,C)$$

$$P(A,B,C,D) = \frac{1}{Z}\,\Phi_1(A,B)\,\Phi_2(C)\,\Phi_3(D)$$

MRF Factorization and pairwise independencies

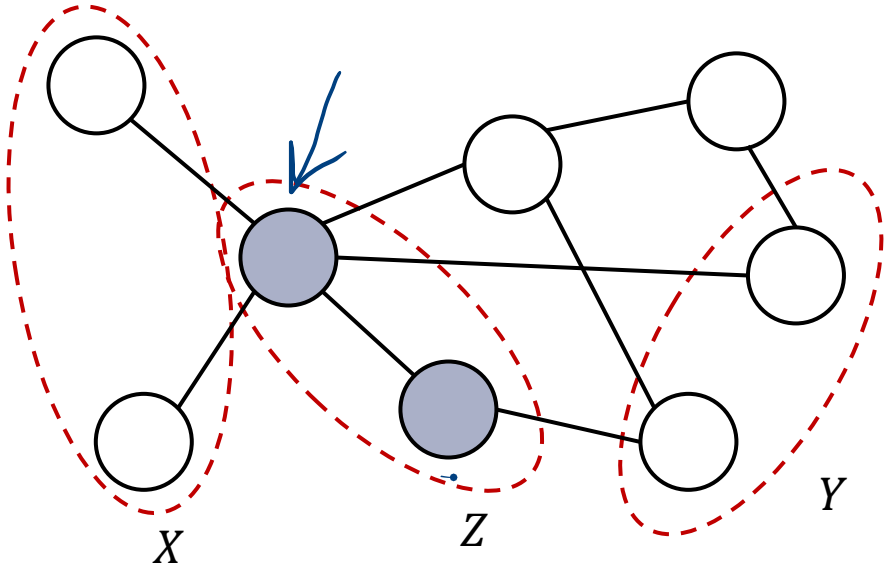
- A distribution $P_{\mathbf{\Phi}}$ with $\mathbf{\Phi} = \{\phi_1(\boldsymbol{D}_1), \ldots, \phi_K(\boldsymbol{D}_K)\}$ **factorizes** over an MRF H if each $\boldsymbol{D}_k$ is a **complete subgraph** of H

- **Potential functions** and **cliques** in the graph completely determine the **joint** distribution.

MRFs: independencies

$X \perp Y | Z$

Separation in the undirected graph:



A path is active given Z if no node in it is in Z

X and Y are separated given Z if there is no active path between X and Y given Z

$$\text{sep}_H(X, Y|Z)$$

- Global independencies for any disjoint sets A, B, C:
  - $A \perp B|C$

If all paths that connect a node in A to a node in B pass through one or more nodes in set C

MRF: independencies

- Determining conditional independencies in undirected models is much easier than in directed ones

- Conditioning in undirected models can only eliminate dependencies while in directed ones observations can create new dependencies (v-structure)

Factorization & Independence

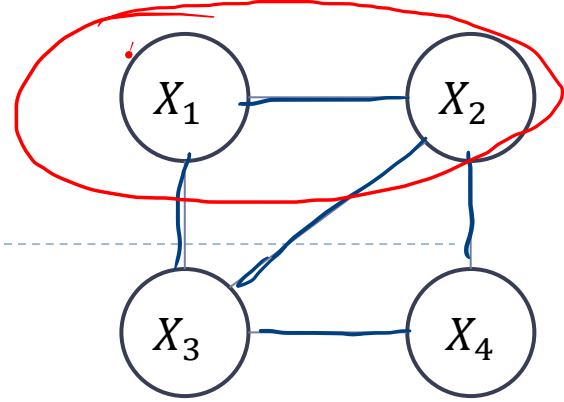
- Factorization ⇒ Independence (soundness of separation criterion)
  - **Theorem:** If P factorizes over H, and $\text{sep}_H(\boldsymbol{X}, \boldsymbol{Y} | \boldsymbol{Z})$ then P satisfies $\boldsymbol{X} \perp \boldsymbol{Y} | \boldsymbol{Z}$ (i.e., H is an I-map of P)
  - $I(H) \subseteq I(P)$

Interpretation of clique potentials

- Potentials cannot all be marginal or conditional distributions

- A positive clique potential can be considered as general compatibility or goodness measure over values of the variables in its scope

Different factorizations



- Maximal cliques:
  - $P_{\mathbf{\Phi}}(X_1, X_2, X_3, X_4) = \frac{1}{Z}\phi_{123}(X_1, X_2, X_3)\phi_{234}(X_2, X_3, X_4)$
  - $Z = \sum_{X_1, X_2, X_3, X_4} \phi_{123}(X_1, X_2, X_3)\phi_{234}(X_2, X_3, X_4)$
- Sub-cliques:
  - $P_{\mathbf{\Phi}'}(X_1, X_2, X_3, X_4) = \frac{1}{Z}\phi_{12}(X_1, X_2)\phi_{23}(X_2, X_3)\phi_{13}(X_1, X_3)\phi_{24}(X_2, X_4)\phi_{34}(X_3, X_4)$
  - $Z = \sum_{X_1, X_2, X_3, X_4} \phi_{12}(X_1, X_2)\phi_{23}(X_2, X_3)\phi_{13}(X_1, X_3)\phi_{24}(X_2, X_4)\phi_{34}(X_3, X_4)$
- Canonical representation
  - $P_{\mathbf{\Phi}'}(X_1, X_2, X_3, X_4) = \frac{1}{Z}\phi_{123}(X_1, X_2, X_3)\phi_{234}(X_2, X_3, X_4)\phi_{12}(X_1, X_2)\phi_{23}(X_2, X_3)\phi_{13}(X_1, X_3) \times \phi_{24}(X_2, X_4)\phi_{34}(X_3, X_4)\phi_1(X_1)\phi_2(X_2)\phi_3(X_3)\phi_4(X_4)$
  - $Z = \sum_{X_1, X_2, X_3, X_4} \phi_{123}(X_1, X_2, X_3)\phi_{234}(X_2, X_3, X_4)\phi_{12}(X_1, X_2)\phi_{23}(X_2, X_3) \times \phi_{13}(X_1, X_3)\phi_{24}(X_2, X_4)\phi_{34}(X_3, X_4)\phi_1(X_1)\phi_2(X_2)\phi_3(X_3)\phi_4(X_4)$

Pairwise MRF

- All of the factors on single variables or pair of variables (X_i, X_j):

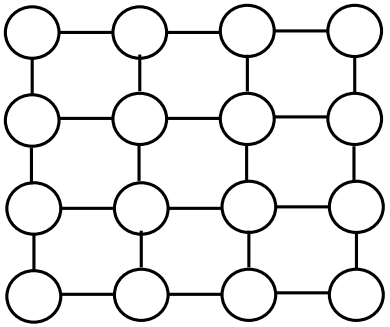
$$P(\boldsymbol{X}) = \frac{1}{Z} \prod_{(X_i, X_j) \in H} \phi_{ij}(X_i, X_j) \prod_i \phi_i(X_i)$$

- Pairwise MRFs are popular (simple special case of general MRFs)

Ising model

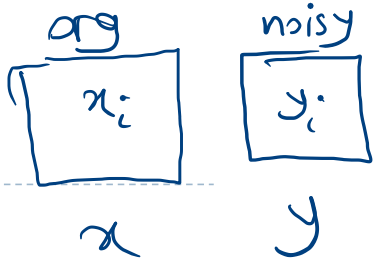
- $X_i \in \{-1,1\}$

$$P(\boldsymbol{x}) = \frac{1}{Z}\exp\left\{\sum_{i}^{n}\theta_i x_i + \sum_{i,j\in E}\alpha_{ij}x_i x_j\right\}$$

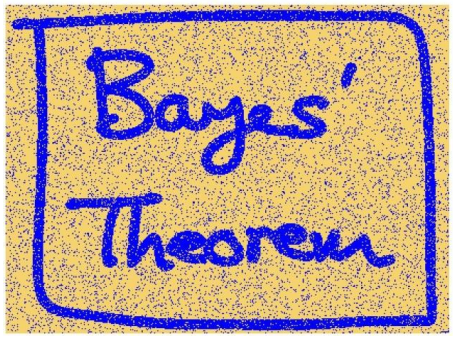


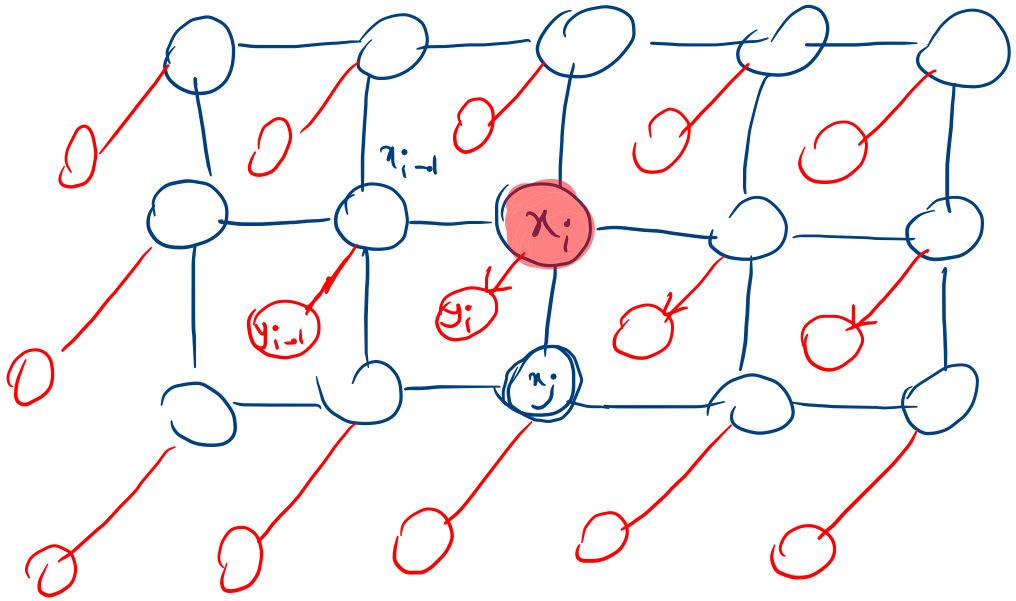
- Grid model
  - Image processing, lattice physics, etc.
  - The states of adjacent nodes are related

Binary Image Denoising



- $y_i \in \{-1,1\}$, array of observed noisy pixels
- $x_i \in \{-1,1\}$, noise free image





$\phi(x_i, x_j)$ $\quad (x_i, x_j) \in E$

$\phi(x_i, y_i)$

$\phi(x_i)$ $\quad \phi(y_i)$

$p = 1/2$ noise $\qquad p = 0.01$

$$P(y=1 \mid x=1) = 1/2$$

$$P(y=-1 \mid x=1) = 1/2$$

$$P(y=+1 \mid x=-1) = 1/2$$

$$P(y=-1 \mid x=-1) = 1/2$$

$$p(y|x) = p(y) = \begin{cases} 1/2 \\ 1/2 \end{cases}$$

$\phi(x_i, x_j) = x_i x_j$ ✗

$= 1 + x_i x_j = 0$ ✗

$= 2 + x_i x_j$ ✓

$= e^{x_i x_j}$

$= e^{-(x_i - x_j)^2}$

$\phi(x_i, y_i) = e^{x_i y_i}$

$p = 0.01$

$p = 0.9$

$e^{-x_i y_i} = e^{x_i(-y_i)}$

$\phi(x_i) = e^{x_i}$

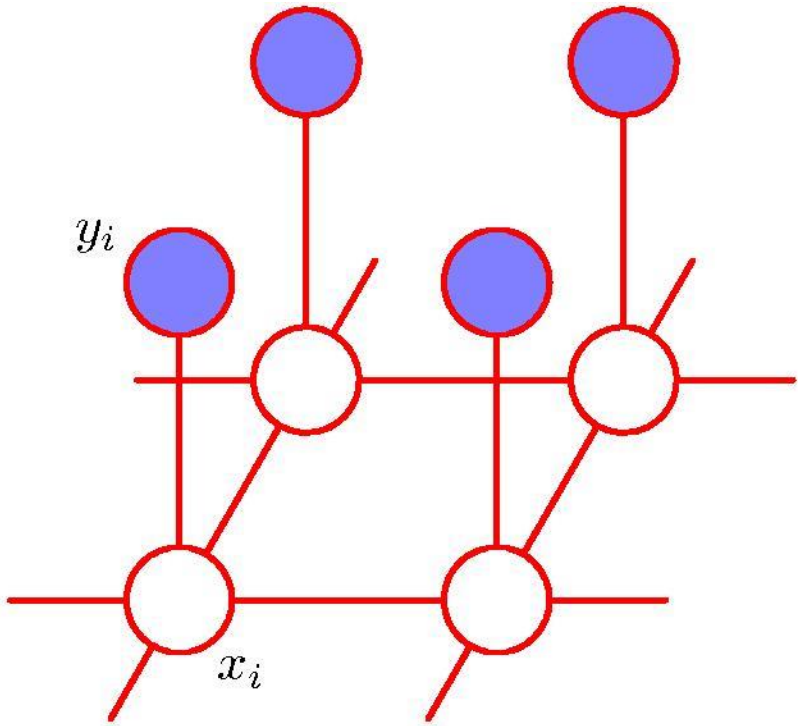
$= e^{-x_i}$

Image denoising

$P(x|y)$

$$\hat{x} = \arg\max_{x} P(x|y)$$

Bayes' Theorem



[Bishop]

$$p(\underline{x}, \underline{y}) = \frac{1}{Z} e^{\alpha \sum_{x_i, x_j \in E} x_i x_j} e^{\beta \sum_i x_i y_i} e^{\gamma \sum_i x_i}$$

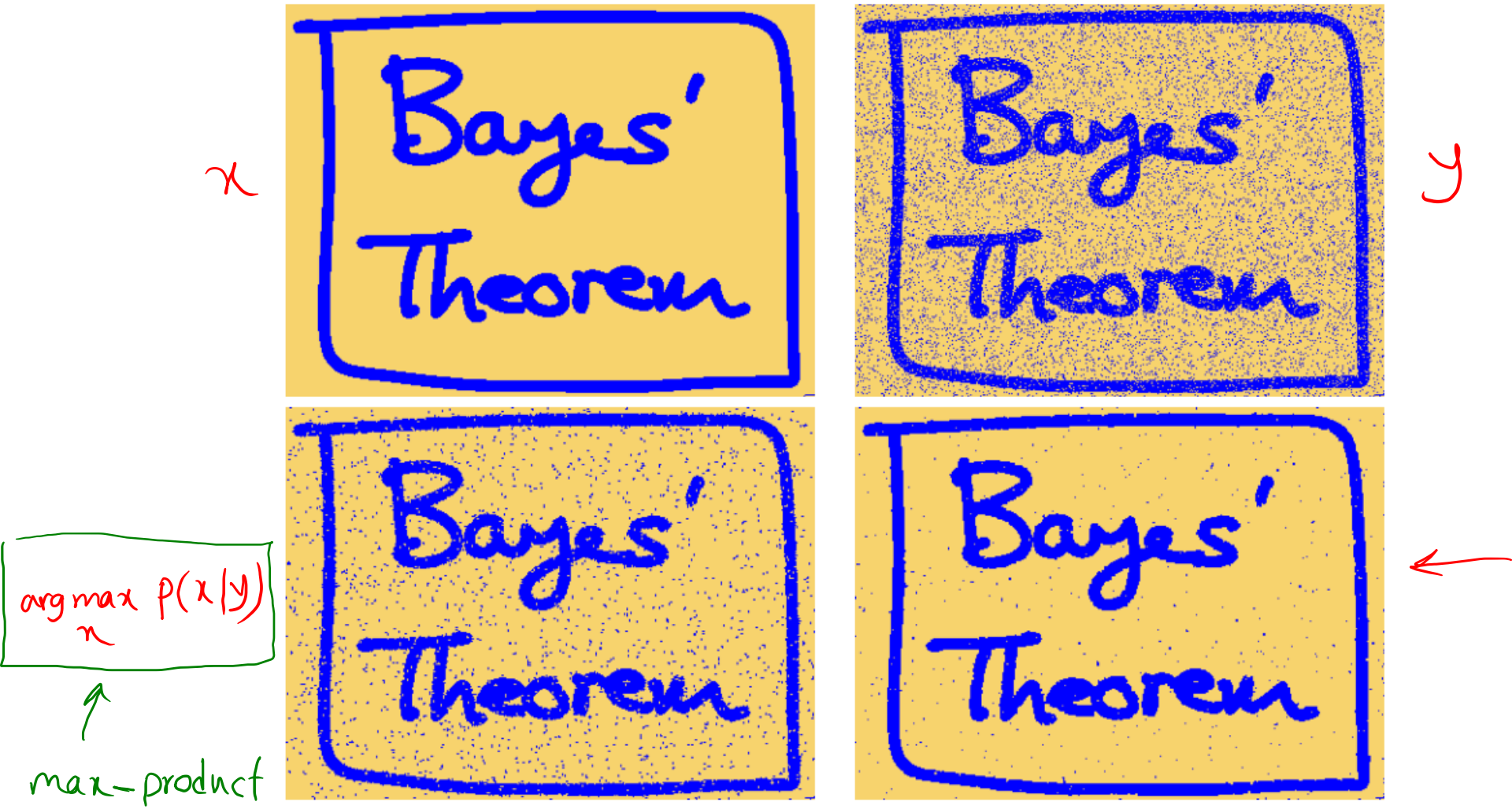


Figure 8.30 Illustration of image de-noising using a Markov random field. The top row shows the original binary image on the left and the corrupted image after randomly changing 10% of the pixels on the right. The bottom row shows the restored images obtained using iterated conditional models (ICM) on the left and using the graph-cut algorithm on the right. ICM produces an image where 96% of the pixels agree with the original image, whereas the corresponding number for graph-cut is 99%.

Bayesian Network

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i \mid Pa(x_i))$$

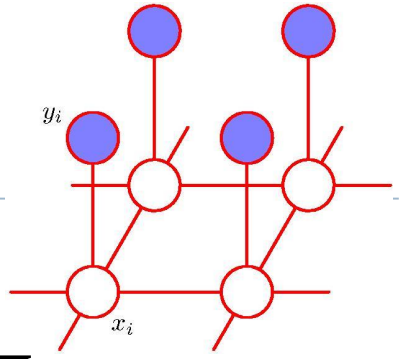
D-separation $\quad I(G) \subseteq I(P)$

Markov Random Fields (MRFs, MN)

$$P(x_1, \ldots, x_n) = \frac{1}{Z} \prod_{i=1}^{K} \Phi_i(D_i)$$

separation

Image denoising



$$P(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{Z} \prod_i \exp\{\gamma x_i y_i\} \prod_i \exp\{\beta x_i\} \prod_{i,j \in H} \exp\{\alpha x_i x_j\}$$

$$= \frac{1}{Z} \exp\left\{ \sum_i \gamma x_i y_i + \sum_i \beta x_i + \sum_{i,j \in H} \alpha x_i x_j \right\}$$

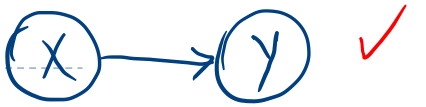
MPA: Most probable assignment of $\boldsymbol{x}$ variables given an evidence $\boldsymbol{y}$

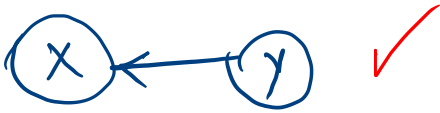
$$\hat{\boldsymbol{x}} = \underset{\boldsymbol{x}}{\operatorname{argmax}}\, P(\boldsymbol{x}|\boldsymbol{y})$$

$$= \arg\max_x \frac{P(x, y)}{P(y)}$$

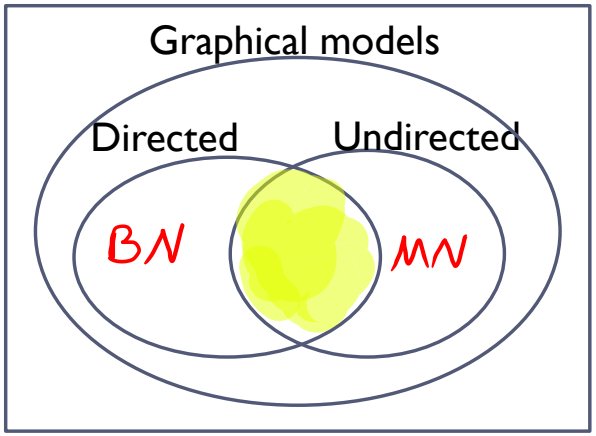
$$= \arg\max_x P(x, y)$$

Perfect map of a distribution





- Not every distribution has a MN perfect map
- Not every distribution has a BN perfect map



Probabilistic models

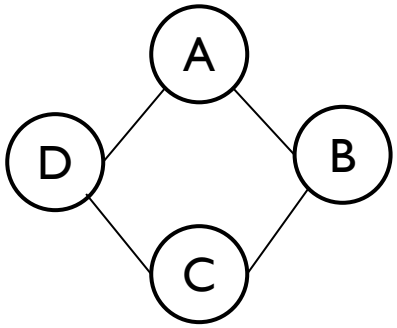
Minimal I-map

- Since we may not find a Markov Network (MN) that is a perfect map of a BN and vice versa, we study the minimal I-map property

- H is a minimal I-map for G if
  - $I(H) \subseteq I(G)$
    - Removal of a single edge in H renders it is not an I-map of G

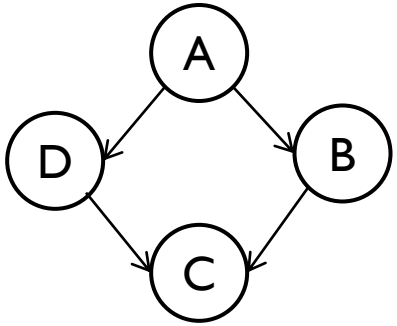
Loop of at least 4 nodes without chord has no equivalent in BNs

- Is there a BN that is a perfect map for this MN?



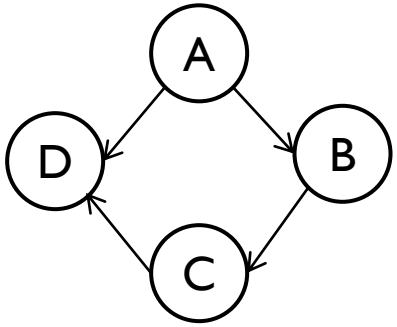
$A \perp C|\{B, D\}$ ✓

$B \perp D|\{A, C\}$ ✓



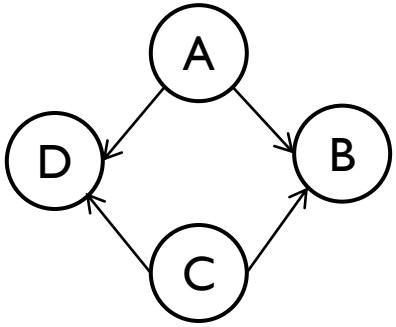
$A \perp C|\{B, D\}$

$B \perp D|\{A, C\}$ ✗



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

$A \perp C|\{B, D\}$ ✗

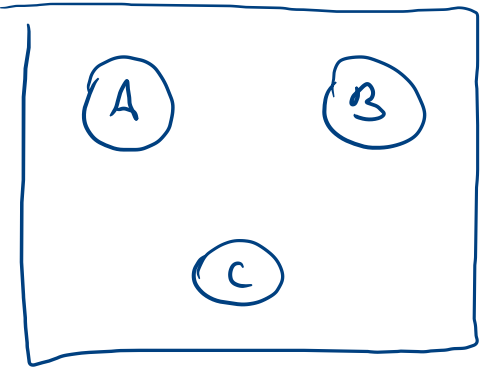


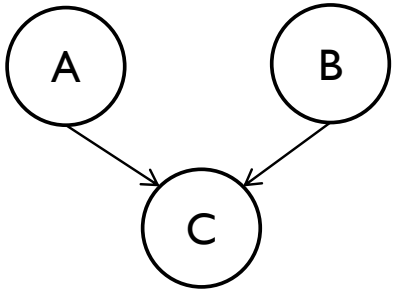
$B \perp D|\{A, C\}$

$A \perp C|\{B, D\}$ ✗

V-structure has no equivalent in MNs

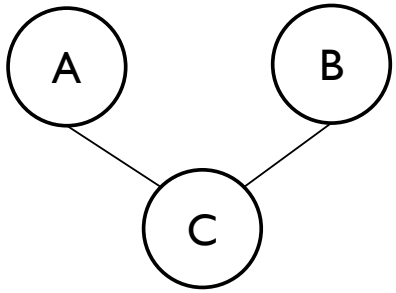
- Is there an MN that is a perfect I-map of this BN?





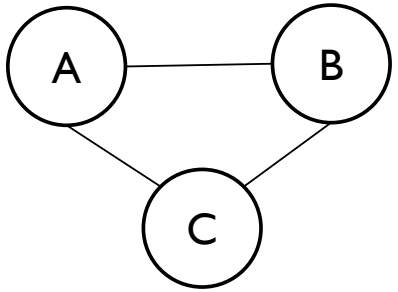
$A \perp B$ ✓

$A \perp B|C$ ✗



$A \perp B$ ✗

$A \perp B|C$



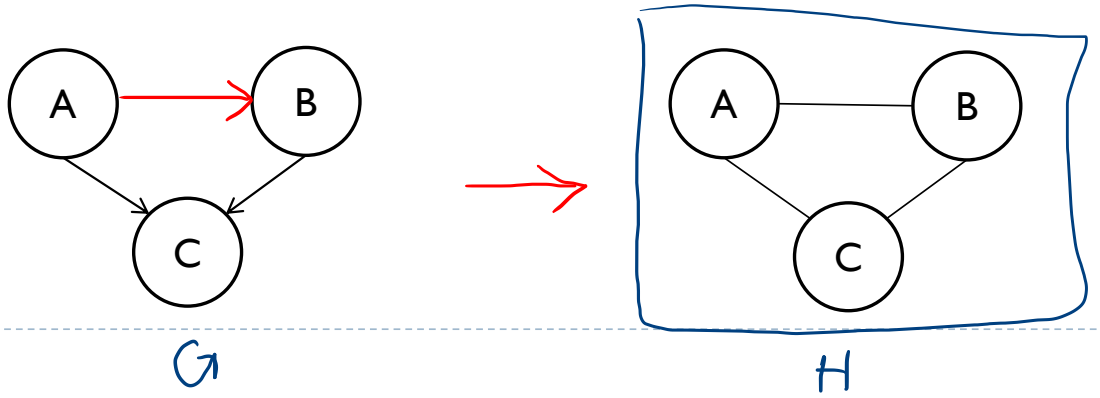
$A \perp B$ ✗

$A \perp B|C$ ✗

Minimal I-maps: from DAGs to MNs

- The **moral graph** $M(G)$ of a DAG G is an undirected graph that contains an undirected edge between X and Y if:
  - there is a directed edge between them in either direction
  - X and Y are parents of the same node

- Moralization turns a node and its parent into a fully connected sub-graph



$I(H) \subseteq I(G)$

Minimal I-maps: from DAGs to MNs

- **Theorem**: The moral graph $M(G)$ of a DAG G is a minimal I-map for G
  - The moral graph loses some independence information
  - But, we cannot remove any edge from it without appearing new independencies that are not in G
    - all independencies in the moral graph are also satisfied in G

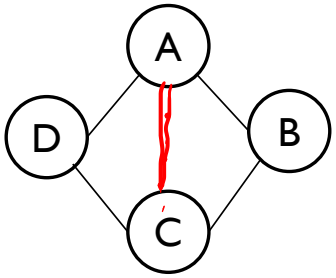
- **Theorem**: If a DAG G is "moral", then its moralized graph $M(G)$ is a perfect I-map of G.

$I(G) \geq I(H)$

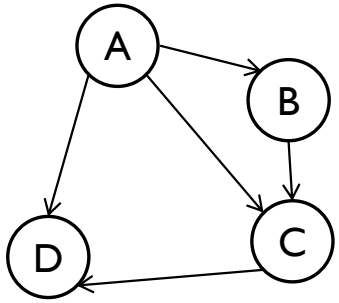
Minimal I-maps: from MNs to DAGs

- **Theorem**: If G is a BN that is minimal I-map for an MN, then G cannot have immoralities.

- **Corollary**: If G is a minimal I-map for an MN then it is **chordal**
  - Any BN that is I-map for an MN must add triangulating edges into the graph



An undirected graph is chordal if any loop with more than three nodes has a chord

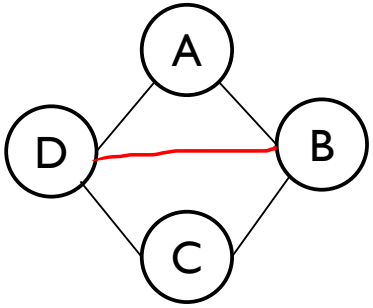


G is a minimal I-map of the left MN

$I(G) \subseteq I(H)$

Perfect I-map

- **Theorem**: Let H be a non-chordal MN. Then there is no BN that is a perfect I-map for H.

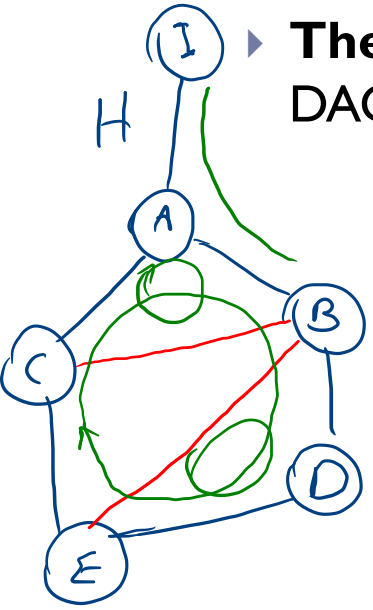


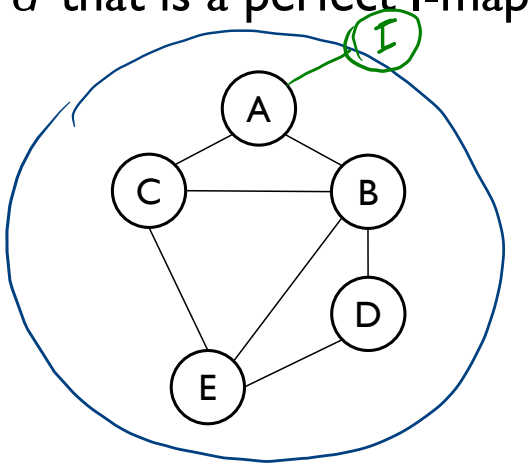
$I(G) = I(P)$

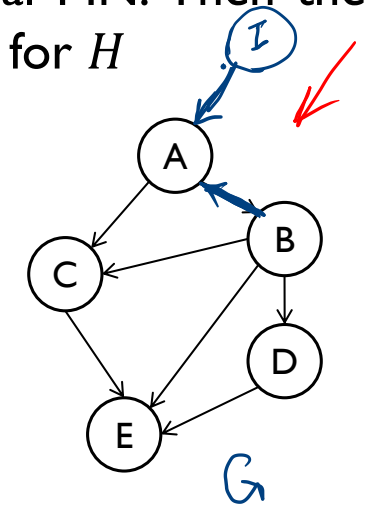
$\Rightarrow$ If the independencies in an MN can be exactly represented via a BN then the MN graph is **chordal**

Perfect I-map

- **Theorem**: Let H be a chordal MN. Then there exists a DAG G that is a perfect I-map for H







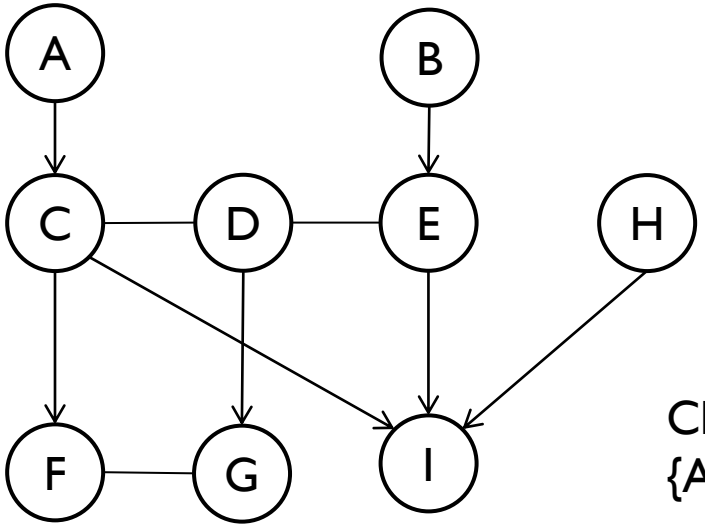
$B \perp I \mid A$ X

$I(G) \subseteq I(H)$

⇒ The independencies in a graph can be represented in both type of models **if and only if** the graph is **chordal**

Partially Directed Acyclic Graphs (PDAGs)

- Superset of both directed and undirected graphs
- PDAGs are also called **chain graphs**



Chain components:
{A}, {B}, {C,D,E},{F,G},{H}, {I}

[Koller, PGM book]

Relationship between BNs and MNs: summary

- Directed and undirected models represent different families of independence assumptions
  - Chordal graphs can be represented in both BNs and MNs
- For inference, we can use a single representation for both types of these models
  - simpler design and analysis of the inference algorithm

Markov Blanket

$MB(X)$ = neighbors of X

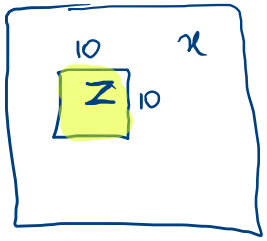
Variational Inference

$p(X_1, X_2, \dots, X_n)$

$p(X_1 \mid X_2, \dots, X_n) = ?$

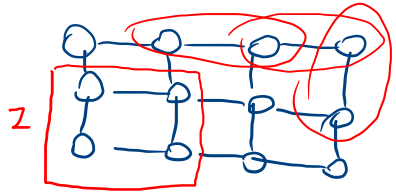
$p(X_1 \mid X_2, X_4) = ?$

Image Inpainting

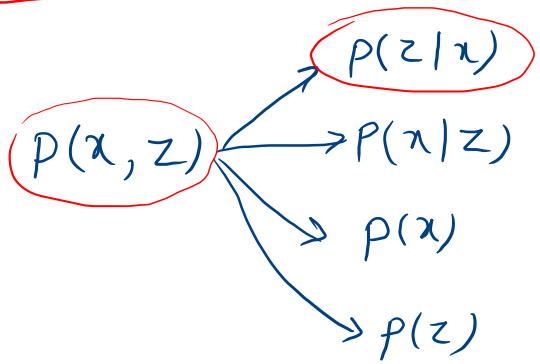


2^{100}

$p(z \mid x)$



$$p(x, z) = \frac{1}{C} \prod_{i=1}^{K} \Phi_i(D_i)$$



$$p(z \mid x) = \frac{p(x, z)}{p(x)} = \frac{p(x, z)}{\sum_z p(x, z)}$$

$S = x \cup z$

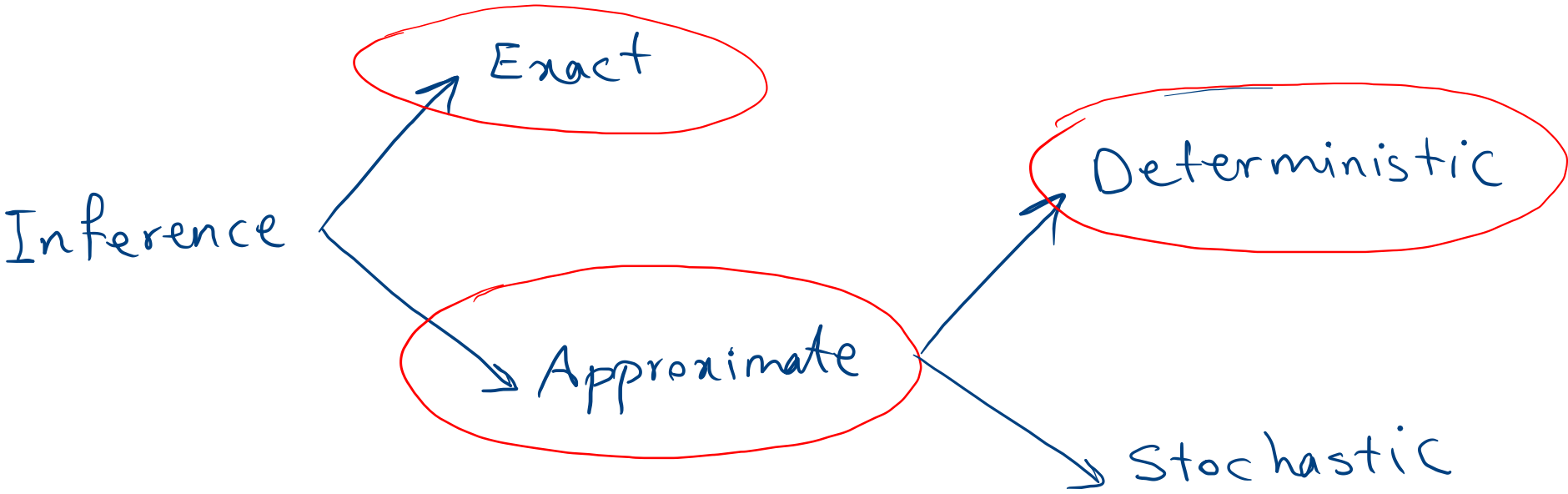
observed

unobserved / Latent / hidden

Inference:

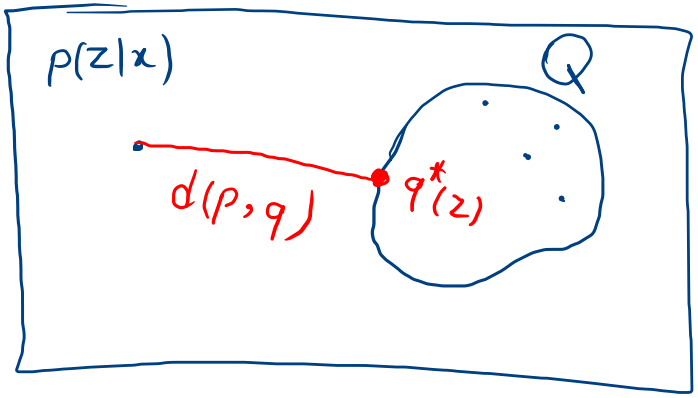
$p(z|x)$

$p(z|x) \simeq q(z)$


Exact
Inference
Deterministic
Approximate
Stochastic

Variational Inference

$p(z|x) \simeq q(z)$ $\quad q(z|x)$



$$q^*(z) = \arg\min_{q(z) \in Q} d\big(q(z), p(z|x)\big)$$

← V. I.

variation calculus

$$\min_x f(x)$$

Functional

variational calculus

$$H(p) = -\sum_x p(x) \log p(x)$$

$p(x)$ $q(x)$

① $d(p,q) = \sum_x |p(x) - q(x)|$

② $d(p,q) = \sum_x (p(x) - q(x))^2$

③ $d(p,q) = \sum_x p(x) |p(x) - q(x)|$

④ $d(p,q) = \sum_x q(x) |p(x) - q(x)|$

x

$p(x) \gg q(x)$

$p(0.2) = 0.9$ $\quad$ $p(0.2) = 0.001$

$q(0.2) = 0.001$ $\quad$ $q(0.2) = 0.9$

$$d(p,q) = \sum_x p(x)\, q(x) |p(x) - q(x)|$$

$p(x) = \epsilon$ $\quad$ $p(x) = 1 - \epsilon$

$q(x) = 2\epsilon$ $\quad$ $q(x) = 1 - 2\epsilon$

$d(x, y)$

① $d(x, y) \geq 0$ ✓ Jensen_inequality

② $x = y \iff d(x, y) = 0$ ✓

③ $d(x, y) = d(y, x)$ ✗

④ $d(x, y) + d(y, z) \geq d(x, z)$ ✗

$$\sum_x \log \frac{p(x)}{q(x)}$$

$p(x) > 0$

$q(x) > 0$

$$KL(p \| q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

KL Divergence

$$KL(q \| p) = \sum_x q(x) \log \frac{q(x)}{p(x)}$$