Georgia Institute of Technology

CSIP
Center for Signal & Image Processing

ECE/ML/CS/ISYE 8803

Probabilistic Graphical Models

Module 3 (Part C):

Undirected Graphical Models

Faramarz Fekri

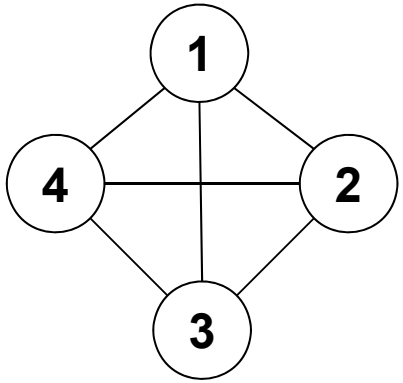
Center for Signal and Information Processing

Overview

- Restricted Boltzmann Machines (RBM)
- Conditional random fields (CRF)
- Gaussian Markov Random Fields (from Chapter 6)
- Partially Directed Graphs

Read Chapter 4 of K&F

Boltzmann Machines



- A fully connected graph with pairwise (edge) potentials on binary-valued nodes (for $x_i \in \{-1,+1\}$ or $x_i \in \{0,1\}$) is called a Boltzmann machine

$$P(x_1, x_2, x_3, x_4) = \frac{1}{Z}\exp\left\{\sum_{ij}\phi_{ij}(x_i, x_j)\right\}$$

$$= \frac{1}{Z}\exp\left\{\sum_{ij}\theta_{ij}x_i x_j + \sum_i \alpha_i x_i + C\right\}$$

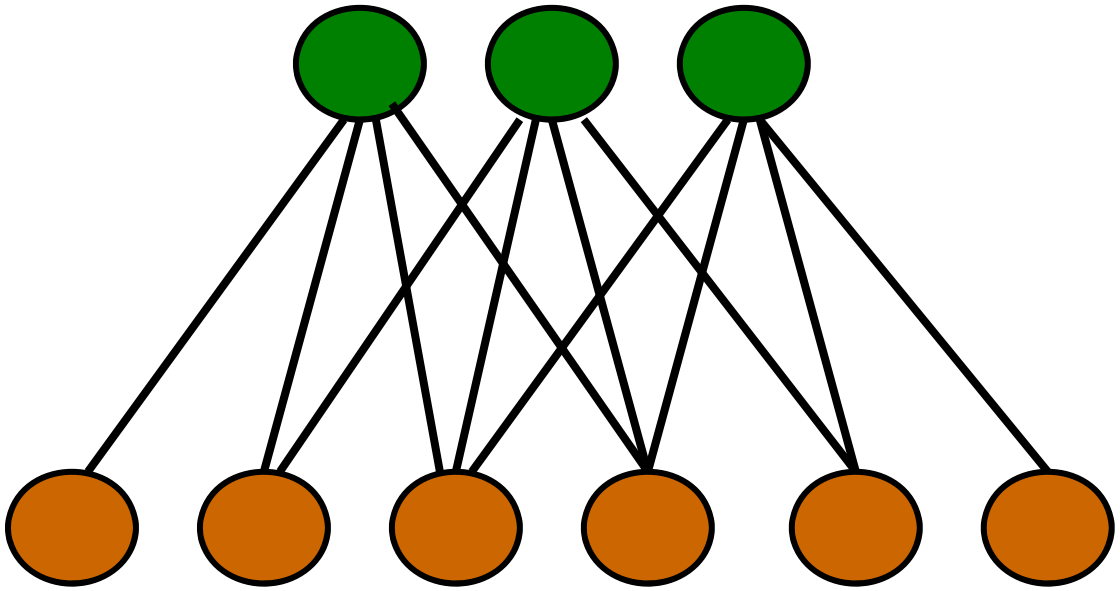
- Hence the overall energy function has the form:

$$H(x) = \sum_{ij}(x_i - \mu)\Theta_{ij}(x_j - \mu) = (x-\mu)^T\Theta(x-\mu)$$

Restricted Boltzmann Machines (RBM)

hidden units

visible units

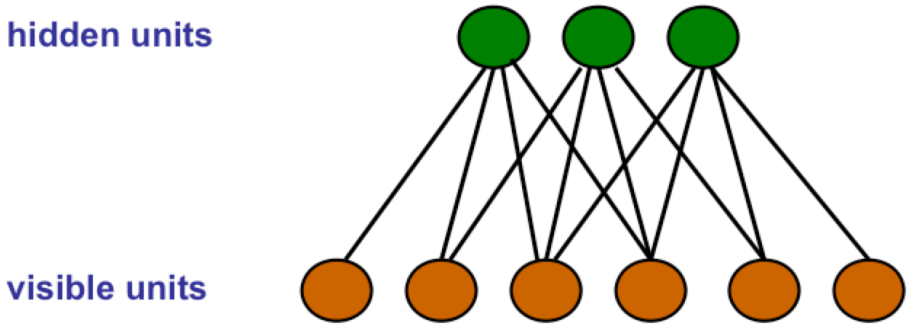


$$p(x,h\,|\,\theta) = \exp\Big\{ \sum_i \theta_i \phi_i(x_i) + \sum_j \theta_j \phi_j(h_j) + \sum_{i,j} \theta_{i,j} \phi_{i,j}(x_i,h_j) - A(\boldsymbol{\theta}) \Big\}$$

Applications:

- dimensionality reduction, classification, regression, collaborative filtering, feature learning and topic modeling

Property of RBM



$$p(\mathbf{v},\mathbf{h}|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})}\exp(-E(\mathbf{v},\mathbf{h};\boldsymbol{\theta}))$$

$$E(\mathbf{v},\mathbf{h};\boldsymbol{\theta}) \triangleq -\sum_{r=1}^{R}\sum_{k=1}^{K} v_r h_k W_{rk} - \sum_{r=1}^{R} v_r b_r - \sum_{k=1}^{K} h_k c_k$$

$$= -(\mathbf{v}^T\mathbf{W}\mathbf{h} + \mathbf{v}^T\mathbf{b} + \mathbf{h}^T\mathbf{c})$$

$$\boldsymbol{\theta} = (\mathbf{W},\mathbf{b},\mathbf{c})$$

$$Z(\boldsymbol{\theta}) = \sum_{\mathbf{v}}\sum_{\mathbf{h}}\exp(-E(\mathbf{v},\mathbf{h};\boldsymbol{\theta}))$$

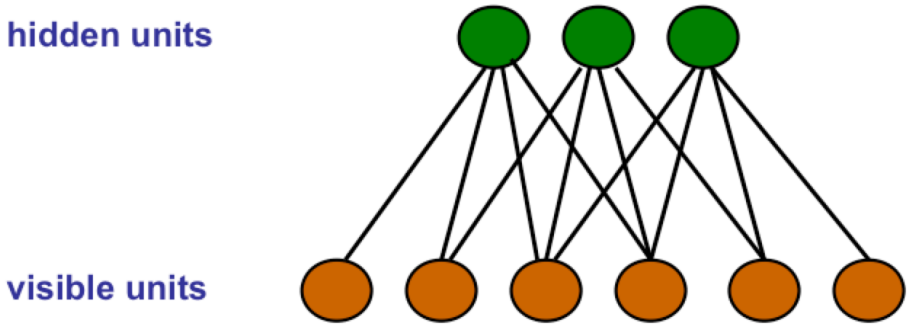
Binary hidden and visible nodes:

Binary RBM

$$p(\mathbf{h}|\mathbf{v},\boldsymbol{\theta}) = \prod_{k=1}^{K} p(h_k|\mathbf{v},\boldsymbol{\theta}) = \prod_{k} \mathrm{Ber}(h_k|\mathrm{sigm}(\mathbf{w}_{:,k}^T\mathbf{v}))$$

$$p(\mathbf{v}|\mathbf{h},\boldsymbol{\theta}) = \prod p(v_r|\mathbf{h},\boldsymbol{\theta}) = \prod \mathrm{Ber}(v_r|\mathrm{sigm}(\mathbf{w}_{r,:}^T\mathbf{h}))$$

Gaussian RBM



$$p(\mathbf{v},\mathbf{h}|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})}\exp(-E(\mathbf{v},\mathbf{h};\boldsymbol{\theta}))$$

$$Z(\boldsymbol{\theta}) = \sum_{\mathbf{v}}\sum_{\mathbf{h}}\exp(-E(\mathbf{v},\mathbf{h};\boldsymbol{\theta}))$$

Gaussian visible nodes (variance =1):

$$E(\mathbf{v},\mathbf{h}|\boldsymbol{\theta}) = -\sum_{r=1}^{R}\sum_{k=1}^{K} W_{rk}h_k v_r - \frac{1}{2}\sum_{r=1}^{R}(v_r - b_r)^2 - \sum_{k=1}^{K} a_k h_k$$

$$p(v_r|\mathbf{h},\boldsymbol{\theta}) = \mathcal{N}(v_r|b_r + \sum_k w_{rk}h_k, 1)$$

Binary hidden node:

$$p(h_k = 1|\mathbf{v},\boldsymbol{\theta}) = \text{sigm}\left(c_k + \sum_r w_{rk}v_r\right)$$

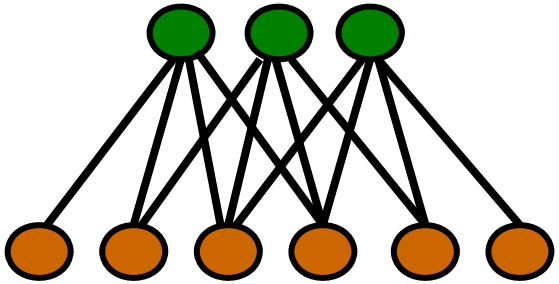
Sampling of RBM (More in Sampling Section)

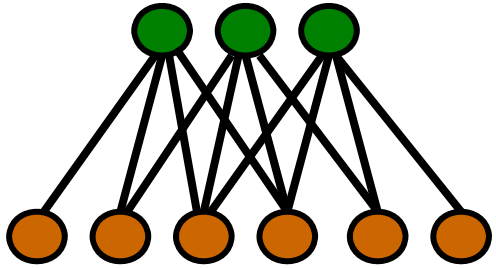
- Factors are marginally *dependent*.

- Factors are conditionally *independent* given observations on the visible nodes.

$$P(\ell \mid \mathbf{w}) = \prod_i P(\ell_i \mid \mathbf{w})$$

- Iterative Gibbs sampling.





$$h \sim p(h \mid x)$$

$$x \sim p(x \mid h)$$

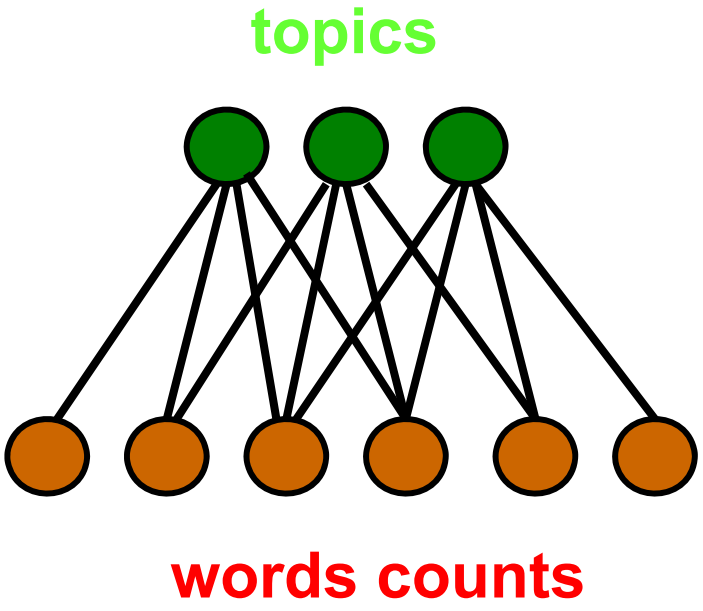
Parameter learning via contrastive divergence (will be visited later)

Example: Topic Modeling

- Topic Modeling:
  - Each topic defines a probability distribution over words (in vocabulary).
  - Each document is represented by a mixture of topics (i.e., mixture of prob. distribution)
  - The mixing proportions of the topics are document specific, but the probability distribution over words, defined by each topic, is the same across all documents.

- Using General Graphical Model is problematic:
  - Exact inference is intractable
  - Lacks distributed representation (i.e., different topics to to give the prob. distribution on "words" collectively.

Example: Topic Modeling via RBM



$h_j = 3$: *topic j has strength 3*

$h_j \in \mathbf{R}, \qquad \langle h_j \rangle = \Sigma_i W_{i,j} x_i$

$x_i = \mathbf{n}$: *word i has count n*

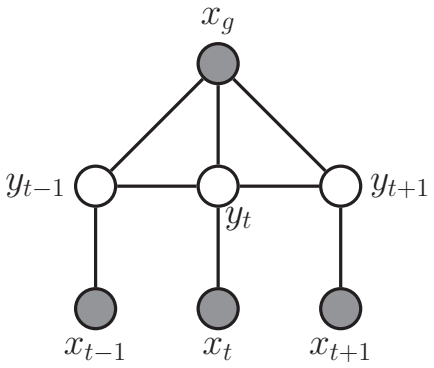
$x_i \in \mathbf{I}$

See the reference paper on RBM for topic modeling:

- R. Salakhutdinov and G. E. Hinton. Replicated Soft- max: an Undirected Topic Model. in NIPS 22, 2009.

Conditional Random Fields (CRFs)

- **Conditional random fields** are undirected graphical models of conditional distributions $p(\mathbf{Y} \mid \mathbf{X})$
  - **Y** is a set of **target variables**
  - **X** is a set of **observed variables**
- We typically show the graphical model using just the **Y** variables
- Potentials are a function of **X** and **Y**
- No node potentials and no edge potentials involving only Xs.
  - Allow arbitrary distribution over X
- Only model relation between $X - Y$ and $Y - Y$



(K. Murphy-page 864)

Formal Definition of CRF

- A CRF is a Markov network on variables $\mathbf{X} \cup \mathbf{Y}$, which specifies the conditional distribution

$$P(\mathbf{y} \mid \mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_{c \in C} \phi_c(\mathbf{x}_c, \mathbf{y}_c)$$

  with partition function

$$Z(\mathbf{x}) = \sum_{\hat{\mathbf{y}}} \prod_{c \in C} \phi_c(\mathbf{x}_c, \hat{\mathbf{y}}_c).$$

- As before, two variables in the graph are connected with an undirected edge if they appear together in the scope of some factor

  The only difference with a standard Markov network is the normalization term – before marginalized over $\mathbf{X}$ and $\mathbf{Y}$, now only over $\mathbf{Y}$

Parameterization of CRF

- Factors may depend on a large number of variables
- We typically parameterize each factor as a log-linear function,

$$\phi_c(\mathbf{x}_c, \mathbf{y}_c) = \exp\{\mathbf{w} \cdot \mathbf{f}_c(\mathbf{x}_c, \mathbf{y}_c)\}$$

- $\mathbf{f}_c(\mathbf{x}_c, \mathbf{y}_c)$ is a feature vector
- $\mathbf{w}$ is a weight vector which is typically learned – we will discuss this extensively in later lectures

- This is without loss of generality: *any* discrete CRF can be parameterized like this (why?)
- Conditional random fields are in the exponential family:

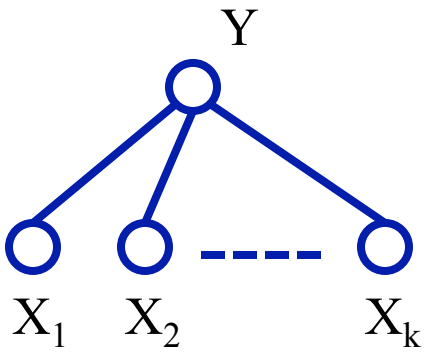
$$\begin{aligned} P(\mathbf{y} \mid \mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_{c \in C} \phi_c(\mathbf{x}, \mathbf{y}_c) &= \exp\left\{ \sum_{c \in C} \mathbf{w}_c \cdot \mathbf{f}_c(\mathbf{x}, \mathbf{y}_c) - \ln Z(\mathbf{w}, \mathbf{x}) \right\} \\ &= \exp\left\{ \mathbf{w} \cdot \mathbf{f}(\mathbf{x}, \mathbf{y}) - \ln Z(\mathbf{w}, \mathbf{x}) \right\}. \end{aligned}$$

Example of CRF

Consider a CRF over the binary-valued variables $\boldsymbol{X} = \{X_1, \ldots, X_k\}$ *and* $\boldsymbol{Y} = \{Y\}$*, and a pairwise potential between* Y *and each* X_i*; this model is sometimes known as a* naive Markov *model, due to its similarity to the naive Bayes model. Assume that the pairwise potentials defined via the following log-linear model*

$$\phi_i(X_i, Y) = \exp\{w_i \mathbb{1}\{X_i = 1, Y = 1\}\}.$$

We also introduce a single-node potential $\phi_0(Y) = \exp\{w_0 \mathbb{1}\{Y = 1\}\}$.



Writing conditional distribution:

$$\tilde{P}(Y = 1 \mid x_1, \ldots, x_k) = \exp\left\{w_0 + \sum_{i=1}^{k} w_i x_i\right\}$$

$$\tilde{P}(Y = 0 \mid x_1, \ldots, x_k) = \exp\{0\} = 1.$$

In this case, we can show

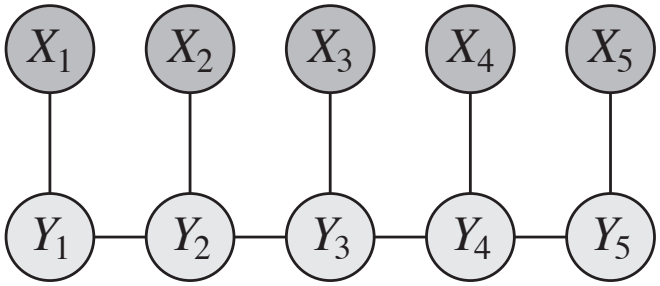
$$P(Y = 1 \mid x_1, \ldots, x_k) = \text{sigmoid}\left(w_0 + \sum_{i=1}^{k} w_i x_i\right) \quad \textit{where} \quad \text{sigmoid}(z) = \frac{e^z}{1 + e^z}$$

Application of CRF in vision

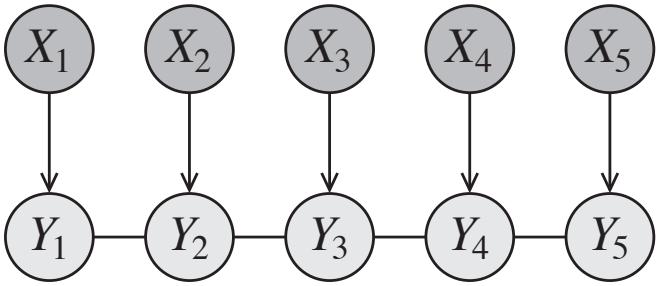
- MRF are popular in computer vision: de-noising, segmentation
- Grid MRF are particularly popular, pixel of an image with 4-pixel neighbors
- **Advantage of CRF over MRF:**
  - No need to spend resources to encode distribution of visible variables. Instead focuses on distribution of labels given data
  - We can make the potentials (or factors) of the model be data-dependent.
    - E.g., in image processing, we may "turn off" the label smoothing between two neighboring nodes **s** and **t** if there is an observed discontinuity in the image intensity between pixels **s** and **t**.

Partially Directed Graphical Models (PDAG or Chain Graphs)

- CRF can be viewed as a special case of a partially directed graphical model:
  - where we have an undirected component over Y, which has the variables in X as parents.



CRF

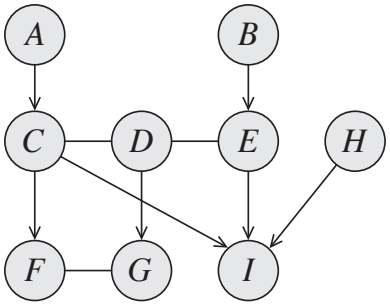


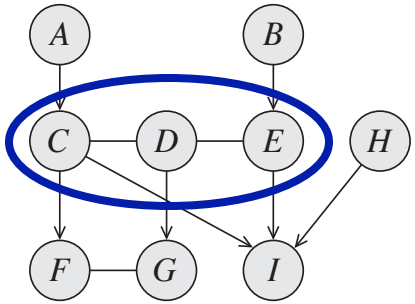
CRF as partially Directed Model

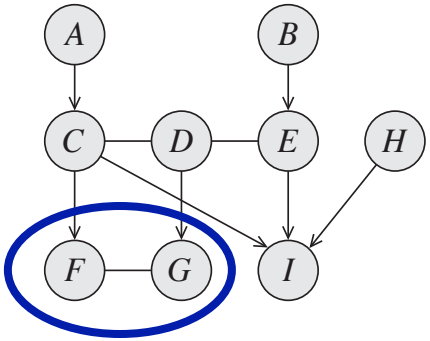
General Case: PDAG (Chain Graphs)

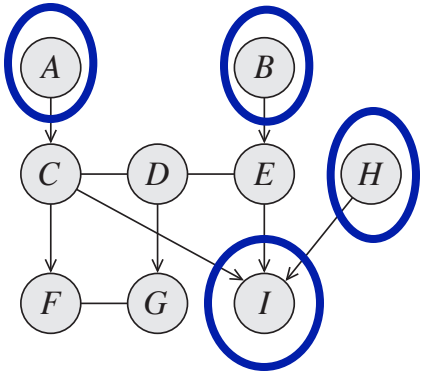
General Framework: partially directed acyclic graph (PDAG):

- Nodes can be disjointly partitioned into several *chain components*.
- An edge between two nodes in the same chain component must be undirected.
- An edge between two nodes in different chain components must be directed.







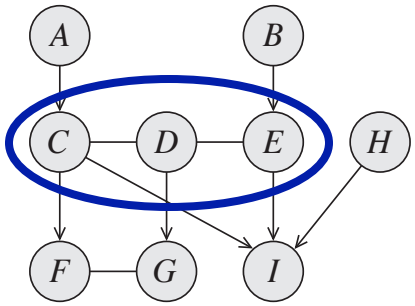


Factorization in PDAG (Chain Graphs)

- *Factorization in partially directed acyclic graph* (PDAG):

Let $\mathcal{K}$ be a PDAG, and $\boldsymbol{K}_1, \ldots, \boldsymbol{K}_\ell$ be its chain components. A chain graph distribution *is defined via a set of factors $\phi_i(\boldsymbol{D}_i)$ $(i = 1, \ldots, m)$, such that each $\boldsymbol{D}_i$ is a complete subgraph in the moralized graph $\mathcal{M}[\mathcal{K}^+[\boldsymbol{D}_i]]$. We associate each factor $\phi_i(\boldsymbol{D}_i)$ with a single chain component $\boldsymbol{K}_j$, such that $\boldsymbol{D}_i \subseteq \boldsymbol{K}_i \cup \mathrm{Pa}_{\boldsymbol{K}_i}$ and define $P(\boldsymbol{K}_i \mid \mathrm{Pa}_{\boldsymbol{K}_i})$ as a CRF with these factors, and with $\boldsymbol{Y}_i = \boldsymbol{K}_i$ and $\boldsymbol{X}_i = \mathrm{Pa}_{\boldsymbol{K}_i}$. We now define*

$$P(\mathcal{X}) = \prod_{i=1}^{\ell} P(\boldsymbol{K}_i \mid \mathrm{Pa}_{\boldsymbol{K}_i}).$$



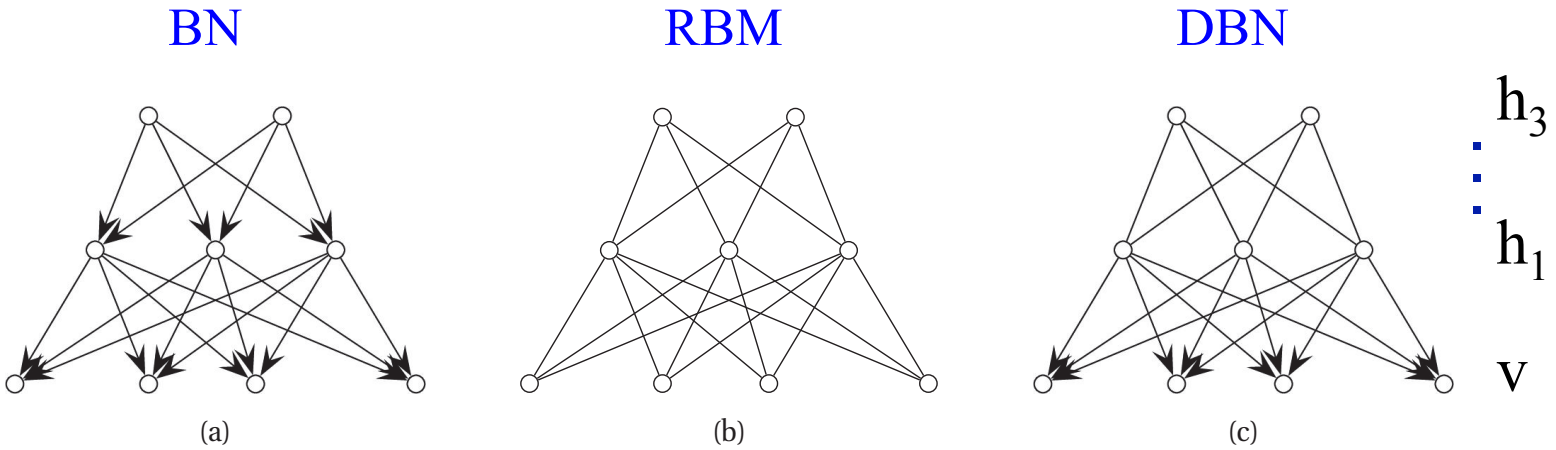
Example:

$P(C, D, E \mid A, B)$ *factorize*

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

See the book, page 149

Deep Belief Networks (DBN) as PDAG



(only v observed)

DBN: a layered model which has directed arrows, except at the top, where there is an undirected bipartite graph.

$$p(\mathbf{h}_1, \mathbf{h}_2, \mathbf{h}_3, \mathbf{v}|\boldsymbol{\theta}) = \prod_i \text{Ber}(v_i|\text{sigm}(\mathbf{h}_1^T \mathbf{w}_{1i}) \prod_j \text{Ber}(h_{1j}|\text{sigm}(\mathbf{h}_2^T \mathbf{w}_{2j})$$

$$\frac{1}{Z(\boldsymbol{\theta})} \exp\left(\sum_{kl} h_{2k} h_{3l} W_{3kl}\right)$$

In above, we assumed all nodes are binary, and all CPDs for directed edges are logistic functions, this is called a **sigmoid belief net**

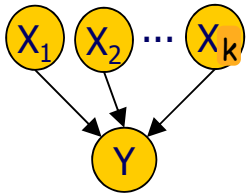
Gaussian Bayesian Networks

Check slides from Module 2 (part B)

- All variables are continuous
- All of the CPDs are linear Gaussians

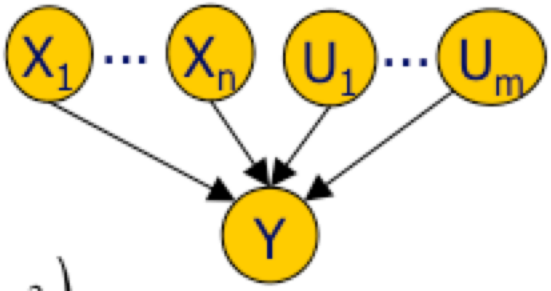
Let Y be a linear Gaussian of its parents $X_1, \ldots, X_k$:

$$p(Y \mid \boldsymbol{x}) = \mathcal{N}\left(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}; \sigma^2\right).$$



Assume that $X_1, \ldots, X_k$ are jointly Gaussian with distribution $\mathcal{N}(\boldsymbol{\mu}; \Sigma)$. Then:

- **Conditional Linear Gaussians**
  - Y continuous variable
  - **X** = {X_1,…,X_n} continuous parents
  - **U** = {U_1,…,U_m} discrete parents
  - $\forall \boldsymbol{u} \in \boldsymbol{U}: \quad P(Y \mid \boldsymbol{u}, \boldsymbol{x}) = N\left(a_{\boldsymbol{u},0} + \sum_{i=1}^{n} a_{\boldsymbol{u},i} \boldsymbol{x}_i; \sigma_{\boldsymbol{u}}^2\right)$



Gaussian Markov Random Fields

- A multivariate Gaussian distribution over $X_1,...X_n$ has
  - nx1 mean vector μ
  - nxn positive definite covariance matrix Σ
    (positive definite: $\forall x \in \Re^n : x^T \sum x > 0$)
  - Joint density function:

$$p(x) = \frac{1}{(2\pi)^{n/2} |\sum|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right]$$

Positive definite Information matrix

$$J = \Sigma^{-1}$$

- *A Gaussian dist. Can be presented by a fully connected MRF with pairwise potentials over continuous variables.*

$$p(\boldsymbol{x}) \propto \exp\left[-\frac{1}{2}\boldsymbol{x}^T J \boldsymbol{x} + (J\boldsymbol{\mu})^T \boldsymbol{x}\right] \qquad \boldsymbol{h} = J\boldsymbol{\mu}$$

- Involve single (quadratic) node potentials and edge potentials correspond to terms:

$$-\frac{1}{2} J_{i,i} x_i^2 + h_i x_i, \qquad -\frac{1}{2}[J_{i,j} x_i x_j + J_{j,i} x_j x_i] = -J_{i,j} x_i x_j$$

Information (Precision) Matrix vs Covariance Matrix

$\Sigma^{-1} =$

1	6	0	0	0
6	2	7	0	0
0	7	3	8	0
0	0	8	4	9
0	0	0	9	5

$\Sigma =$

0.10	0.15	-0.13	-0.08	0.15
0.15	-0.03	0.02	0.01	-0.03
-0.13	0.02	0.10	0.07	-0.12
-0.08	0.01	0.07	-0.04	0.07
0.15	-0.03	-0.12	0.07	0.08

$\Sigma_{15}^{-1} = 0 \Leftrightarrow X_1 \perp X_5 \mid TheRest$

~~$X_1 \perp X_5 \Leftrightarrow \Sigma_{15} = 0$~~