Alternative Parameterizations of Markov Networks

Sargur Srihari
srihari@cedar.buffalo.edu

Topics

- Three types of parameterization
  1. Gibbs Parameterization shortcomings
  2. Factor Graphs
  3. Log-linear Models with Energy functions
     - Log-linear with Features
     - Ising, Boltzmann
- Overparameterization
  - Canonical Parameterization
  - Eliminating Redundancy

Gibbs Parameterization

- A distribution P_Φ is a Gibbs distribution parameterized by a set of factors

$$\Phi = \{\phi_1(D_1),..,\phi_K(D_K)\}$$

  - If it is defined as follows

$$P_\Phi(X_1,..X_n) = \frac{1}{Z}\tilde{P}(X_1,..X_n)$$

where

$$\tilde{P}(X_1,..X_n) = \prod_{i=1}^{m}\phi_i(D_i)$$

is an unnomalized measure and

$$Z = \sum_{X_1,..X_n}\tilde{P}(X_1,..X_n)$$

D_i are sets of random variables

A factor ϕ is a function from $Val(\boldsymbol{D})$ to R where Val is the set of values that $\boldsymbol{D}$ can take

Factor returns a "potential"

The factors do not necessarily represent the marginal distributions $p(D_i)$ of the variables in their scopes

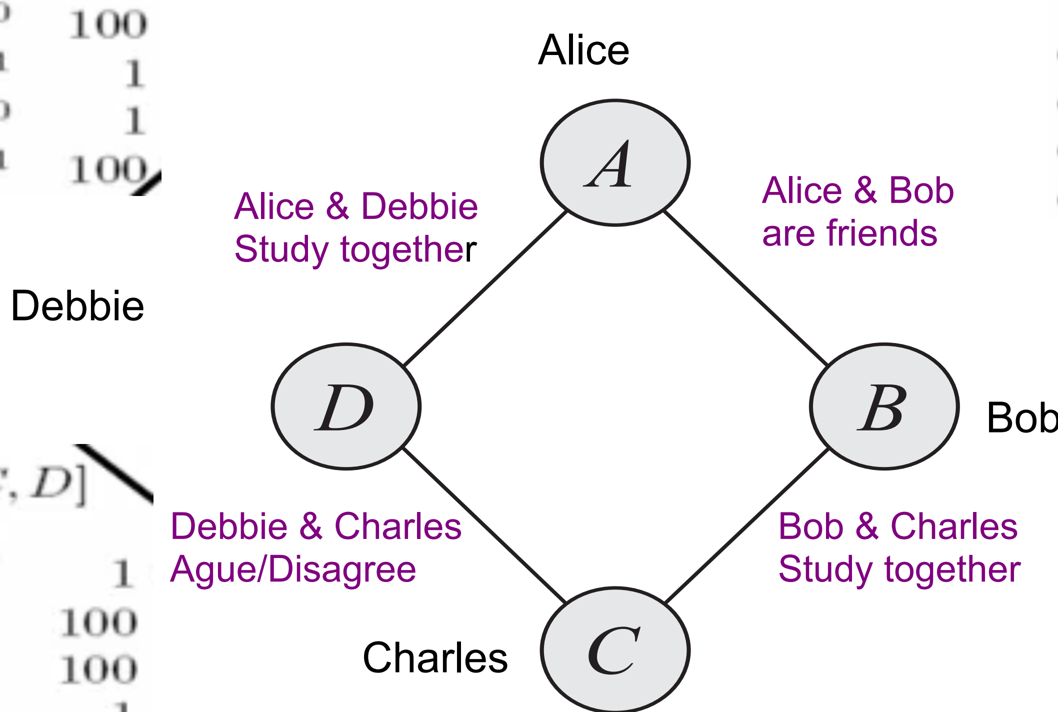
Gibbs Parameters with Pairwise Factors

$\phi_4[D, A]$

d^0	a^0	100
d^0	a^1	1
d^1	a^0	1
d^1	a^1	100

$\phi_1[A, B]$

a^0	b^0	30
a^0	b^1	5
a^1	b^0	1
a^1	b^1	10



$\phi_3[C, D]$

c^0	d^0	1
c^0	d^1	100
c^1	d^0	100
c^1	d^1	1

$\phi_2[B, C]$

b^0	c^0	100
b^0	c^1	1
b^1	c^0	1
b^1	c^1	100

$$P(a,b,c,d) = \frac{1}{Z}\phi_1(a,b)\cdot\phi_2(b,c)\cdot\phi_3(c,d)\cdot\phi_4(d,a)$$

where

$$Z = \sum_{a,b,c,d}\phi_1(a,b)\cdot\phi_2(b,c)\cdot\phi_3(c,d)\cdot\phi_4(d,a)$$

Note that Factors are Non-negative

Shortcoming of Gibbs Parameterization

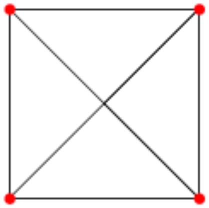
- Network structure doesn’t reveal parameterization
- Cannot tell whether the factors are maximal cliques or subsets
- Example next

Two Gibbs parameterizations, same MN structure

- Gibbs distribution P over fully connected graph
  1. Clique potential parameterization
     - Entire graph is a clique

$$P(a,b,c,d)=\frac{1}{Z}\phi(a,b,c,d) \text{ where } Z=\sum_{a,b,c,d}\phi\,(a,b,c,d)$$

     - No of Parameters
       - » Exponential in no. of variables: 2^n-1



Completely connected graph with four binary variables

  2. Pairwise parameterization
     - A factor for each pair of variables X,Y ε χ

$$P(a,b,c,d)=\frac{1}{Z}\phi_1(a,b)\cdot\phi_2(b,c)\cdot\phi_3(c,d)\cdot\phi_4(d,a)\cdot\phi_5(a,c)\cdot\phi_6(b,d) \text{ where } Z=\sum_{a,b,c,d}\phi_1(a,b)\cdot\phi_2(b,c)\cdot\phi_3(c,d)\cdot\phi_4(d,a)\cdot\phi_5(a,c)\cdot\phi_6(b,d)$$

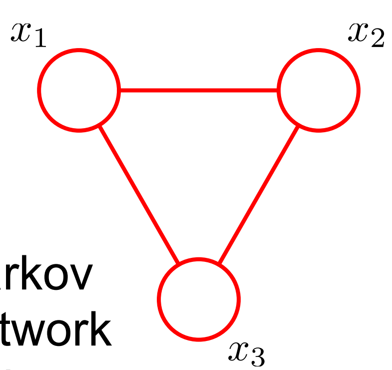
     - Quadratic no of parameters: $4 \times {}^nC_2$
- Independencies are same in both
  - But significant difference in no of parameters

Factor Graphs

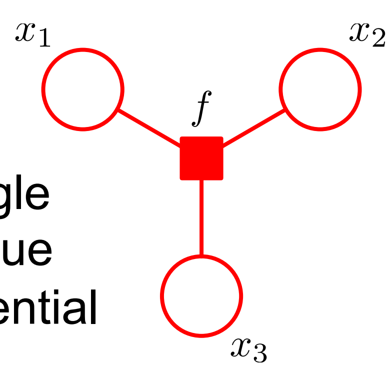
- Markov network structure does not reveal all structure in a Gibbs parameterization
  - Cannot tell from graph whether factors involve maximal cliques or their subsets
  - Factor graph makes parameterization explicit

Factor Graph

- Undirected graph with two types of nodes
  - Variable nodes denoted as ovals
  - Factor nodes denoted as squares
- Contains edges only between variable nodes and factor nodes

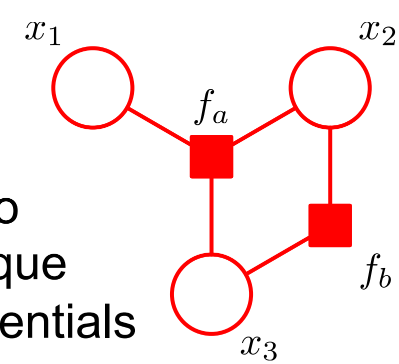


Markov Network of three variables



Single Clique potential

$$P(x_1,x_2,x_3)=\frac{1}{Z}f(x_1,x_2,x_3)$$
$$where\ Z=\sum_{x1,x2,x3} f(x_1,x_2,x_3)$$



Two Clique potentials

$$P(x_1,x_2,x_3)=\frac{1}{Z}f_a(x_1,x_2,x_3)f_b(x_2,x_3)$$
$$where\ Z=\sum_{x1,x2,x3} f_a(x_1,x_2,x_3)f_b(x_2,x_3)$$

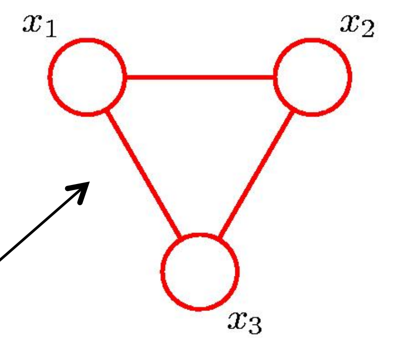
Parameterization of Factor Graphs

- MN parameterized by a set of factors
- Each factor node V_ϕ is associated
  - with only one factor ϕ
  - whose scope is the set of variables that are neighbors of V_ϕ

A distribution P factorizes over Factor graph $\mathcal{F}$ if it can be represented as a set of factors in this form

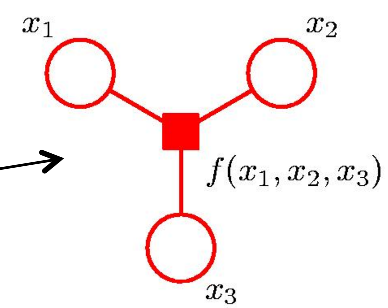
Multiple factor graphs for same graph

- Factor graphs are specific about factorization
- A fully connected undirected graph



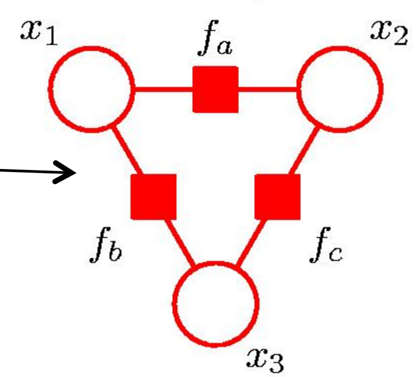
- Joint distribution in two forms
  - In general form

$$p(x)=f(x_1,x_2,x_3)$$

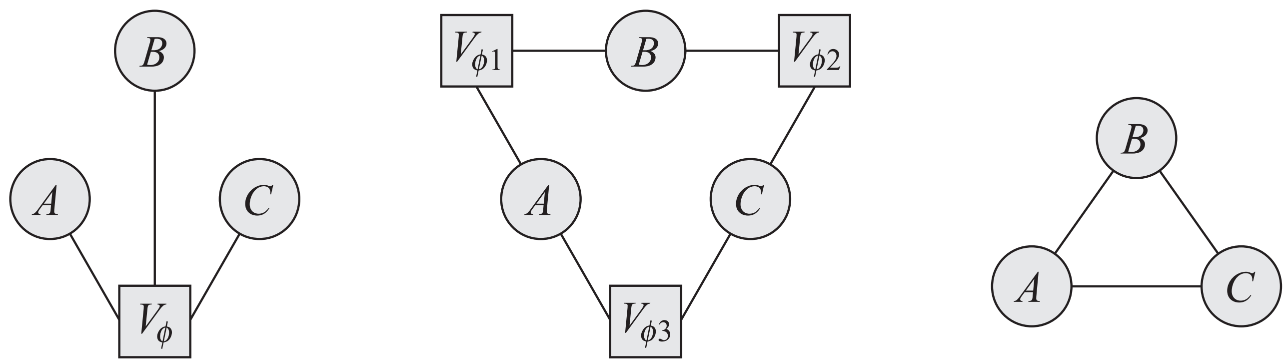


  - As a specific factorization

$$p(x)=f_a(x_1,x_2)f_b(x_1,x_3)f_c(x_2,x_3)$$



Factor graphs for same network



(a) (b) (c)

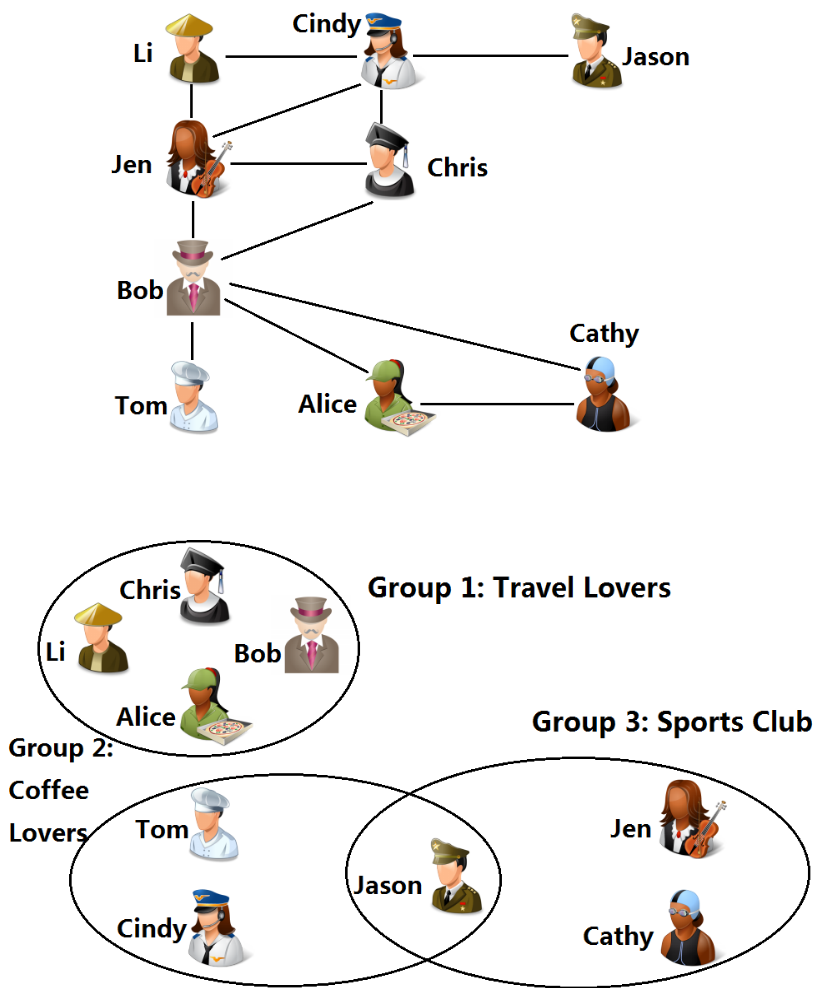
Single factor over all variables

Three pairwise factors

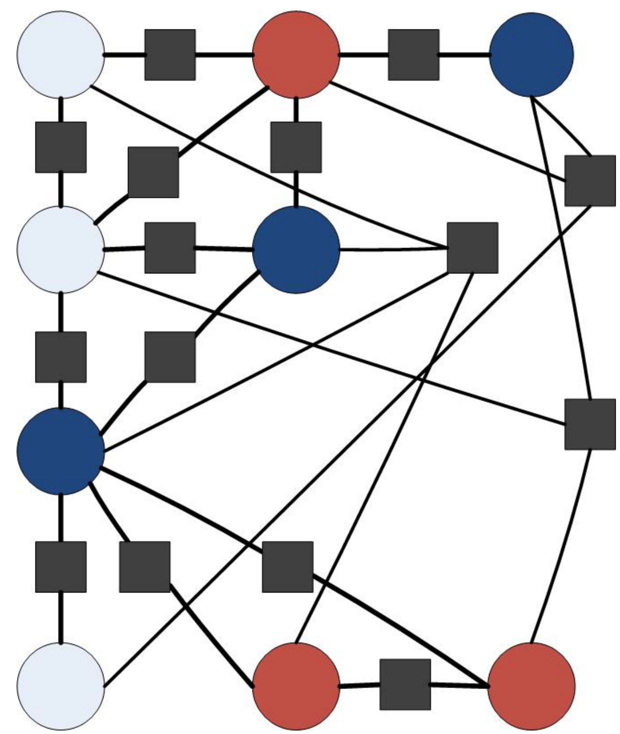
Induced Markov network for both is a clique over A,B,C

- Factor graphs (a) and (b) imply the same Markov network (c)
- Factor graphs make explicit the difference in factorization

Social Network Example

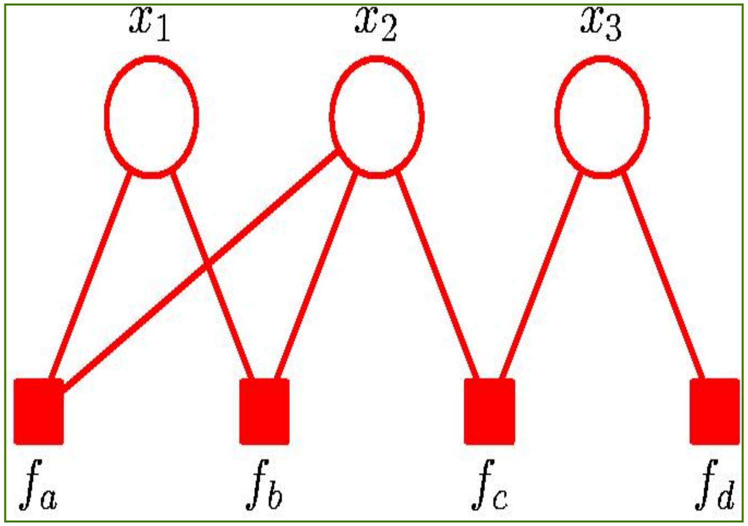


Factor graph with pairwise and Higher-order factors



Factor graphs properties

- They are bipartite since
  1. Two types of nodes
  2. All links go between nodes of opposite type
- Representable as two rows of nodes
  - Variables on top
  - Factor nodes at bottom
- Other intuitive representations used
  - When derived from directed/ undirected graphs

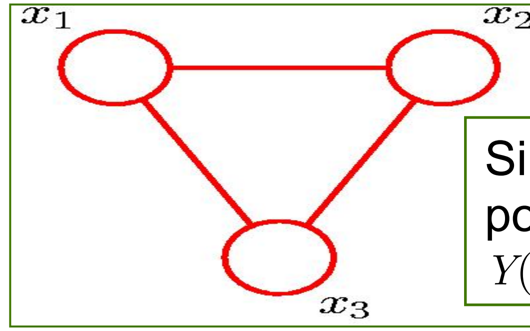


Deriving factor graphs from Graphical Models

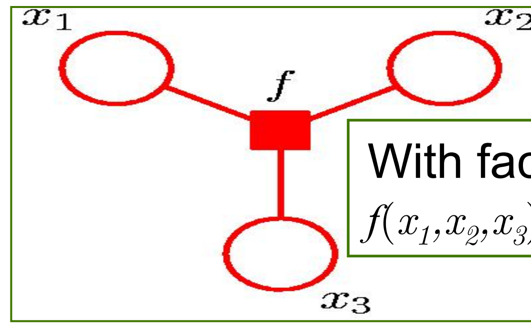
- Undirected Graph (MN)
- Directed Graph (BN)

Conversion of MN to Factor Graph

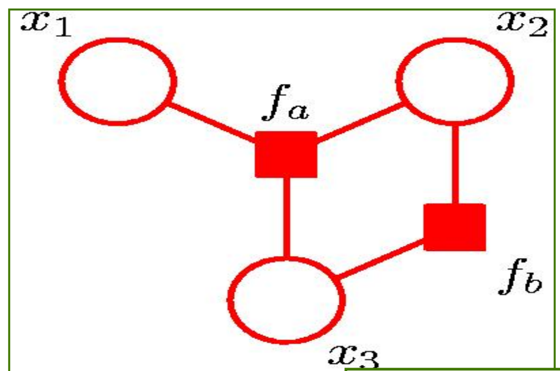
- Steps in converting distribution expressed as undirected graph:
  1. Create variable nodes corresponding to nodes in original
  2. Create factor nodes for maximal cliques x_s
  3. Factors $f_s(\mathrm{x}_s)$ set equal to clique potentials
- Several different factor graphs possible from same distribution



Single clique potential
$Y(x_1,x_2,x_3)$



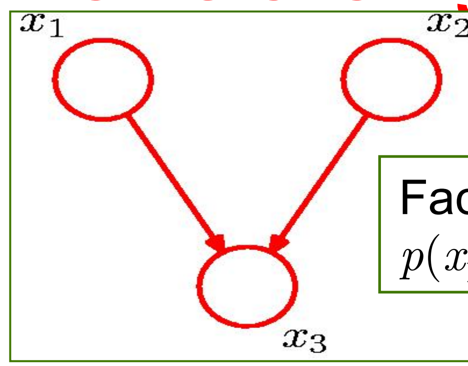
With factors
$f(x_1,x_2,x_3) = Y(x_1,\ x_2,\ x_3)$



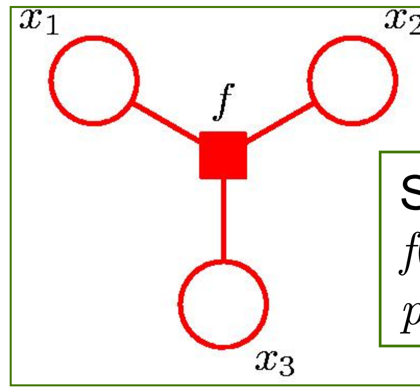
With factors
$f_a(x_1,x_2,x_3)f_b(x_2,x_3) = Y(x_1,x_2,x_3)$

Conversion of BN to factor graph

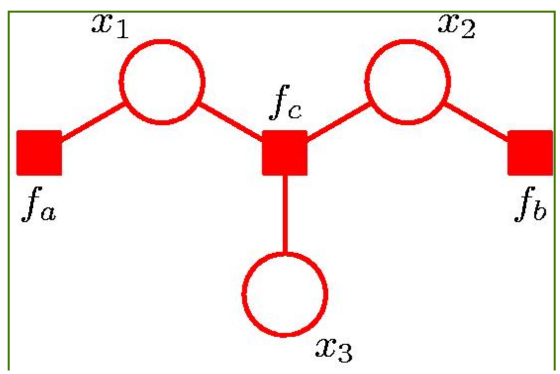
- Steps
  1. Variable nodes correspond to nodes in factor graph
  2. Create factor nodes corresponding to conditional distributions
     - Multiple factor graphs possible from same graph



Factorization
$p(x_1)p(x_2)p(x_3|x_1,x_2)$



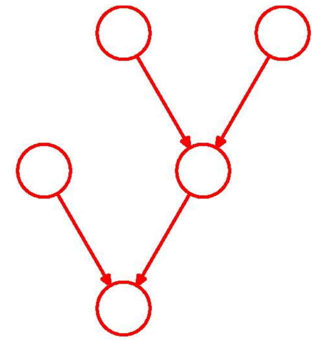
Single Factor
$f(x_1,x_2,x_3)=$
$p(x_1)p(x_2)p(x_3|x_1,x_2)$



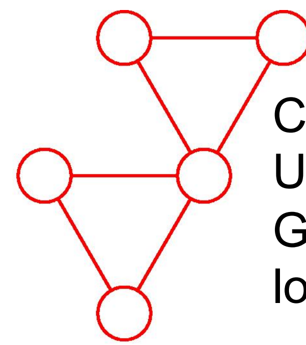
With three Factors
$f_a(x_1)=p(x_1)$
$f_b(x_2)=p(x_2)$
$f_c(x_1,x_2,x_3)=p(x_3|x_1,x_2)$

Tree to Factor Graph

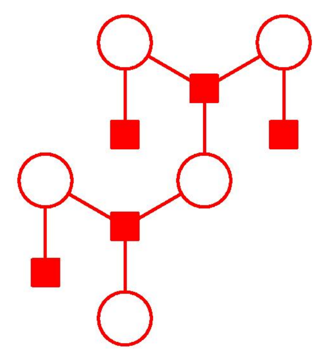
- Conversion of directed or undirected tree to factor graph is a tree
  - No loops
  - Only one path between 2 nodes
- In the case of a directed polytree
  - Conversion to undirected graph has loops due to moralization
  - Conversion again to factor graph results in a tree



Directed polytree



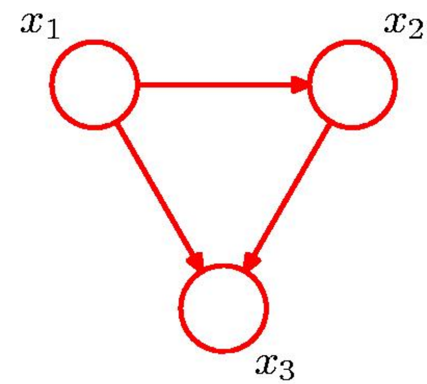
Converted to Undirected Graph with loops

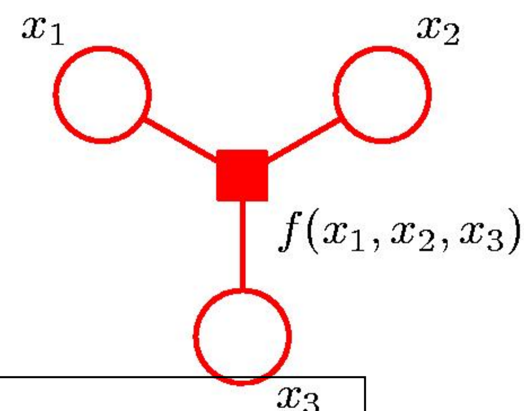


Factor Graphs

Removal of local cycles

- Local cycles in a directed graph having links connecting parents
- Can be removed on conversion to factor graph
  - By defining a factor function





Factor Graph with tree structure

$$f(x_1, x_2, x_3) = p(x_1)\ p(x_2 \mid x_1)\ p(x_3 \mid x_1, x_2)$$

Log-linear Models

- As in Gibbs parameterization,
  - Factor graphs still encode factors as tables over its scope

$$P(x_1,x_2,x_3)=\frac{1}{Z}f_a(x_1,x_2,x_3)f_b(x_2,x_3)$$
$$where\ Z=\sum_{x1,x2,x3} f_a(x_1,x_2,x_3)f_b(x_2,x_3)$$

- Factors can also exhibit Context-specific structure (as in BNs)
- Patterns more readily seen in log-space

Conversion to log-space

- A factor $\phi(D)$ is a function from $Val(D)$ to R
  - where Val is the set of values that D can take
- Rewrite factor $\phi(D)$ as

$$\phi(D) = \exp(-\varepsilon(D))$$

  - Where $\varepsilon(D)$ is the *energy* function defined as

$$\varepsilon(D) = -\ln \phi(D)$$

    - Note that if $\ln a = -b$ then $a=\exp(-b)$
    - If we have $a=\phi(D)=\exp(-\varepsilon(D))$ then $b=-\ln a=\varepsilon(D)$
- Thus factor value $\phi(D)$, a probability, is negative exponential of energy value $\varepsilon(D)$

Energy: Terminology of Physics

- Higher energy states have lower probability
- D is a set of atoms with their values being states and $\varepsilon(D)$ is its energy, a scalar
- Probability $\phi(D)$ of a physical state depends inversely on its energy

$$\phi(D) = \exp(-\varepsilon(D)) = \frac{1}{\exp(\varepsilon(D))}$$

- "Log linear" is term used in field of statistics for logarithms of cell frequencies

$$\varepsilon(D) = -\ln\phi(D)$$

Probability in logarithmic representation

$$P_\Phi(X_1,..X_n) = \frac{1}{Z}\tilde{P}(X_1,..X_n)$$

where

$$\tilde{P}(X_1,..X_n) = \prod_{i=1}^{m} \phi_i(D_i)$$

is an unnomalized measure and

$$Z = \sum_{X_1,..X_n} \tilde{P}(X_1,..X_n)$$

$$P(X_1,...,X_n) \;\; \alpha \;\; \exp\left[-\sum_{i=1}^{m} \varepsilon_i(D_i)\right]$$

Since

$$\phi_i(D_i) = \exp(-\varepsilon_i(D_i))$$

where $\varepsilon_i(D_i) = -\ln(\phi_i(D_i))$

- Taking logarithm requires that $\phi(D)$ be positive
  - Note that probability is positive
- Log-linear parameters $\varepsilon(D)$ can be any value along the real line
  - Not just non-negative as with factors
- Any Markov network parameterized using positive factors can be converted into a log-linear representation

Partition Function in Physics

$$P(X_1,..,X_n) = \frac{1}{Z} \exp\left[-\sum_{i=1}^{m} \varepsilon_i(D_i)\right]$$

where

$$Z = \sum_{X_1,..X_n} \exp\left[-\sum_{i=1}^{m} \varepsilon_i(D_i)\right]$$

- Z describes the statistical properties of a system in thermodynamic equilibrium
  - They are functions of thermodynamic state variables such as temperature and volume
  - it encodes how probabilities are partitioned among different microstates, based on their individual energies
  - Z for *Zustandssumme*, "sum over states"

Log-linear Parameterization

- To convert factors in Gibbs parameterization to log-linear form:
  - Take negative natural logarithm of each potential
    - Requires potential to be positive (to take logarithm)

$\phi_1[A, B]$

a^0	b^0	30
a^0	b^1	5
a^1	b^0	1
a^1	b^1	10

$$\xrightarrow{-\ln 30 = -3.4}$$

$\epsilon_1(A, B)$

a^0	b^0	-3.4
a^0	b^1	-1.61
a^1	b^0	0
a^1	b^1	-2.3

$$\varepsilon(D) = -\ln \phi(D)$$

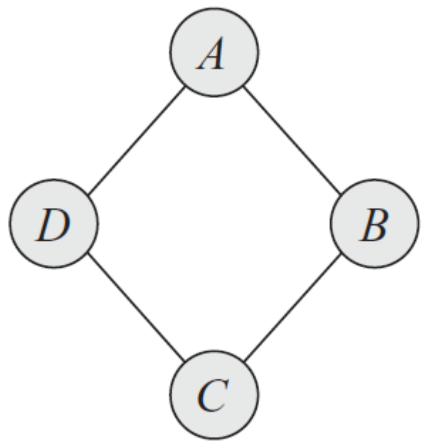
Thus

$$\phi(D) = \exp(-\varepsilon(D))$$

Energy is
Negative log probability

Example

$$P(X_1,..,X_n) \;\alpha\; \exp\left[-\sum_{i=1}^{m} \varepsilon_i(D_i)\right]$$



$$P(A,B,C,D) \;\alpha\; \exp\left[-\varepsilon_1(A,B)-\varepsilon_2(B,C)-\varepsilon_3(C,D)-\varepsilon_4(D,A)\right]$$

Partition function Z is sum of RHS over all values of A,B,C,D

Product of Factors becomes sum of exponentials , e.g.,

$$\phi(A,B)=\exp(-\varepsilon(A,B))$$

From Potentials to Energy Functions

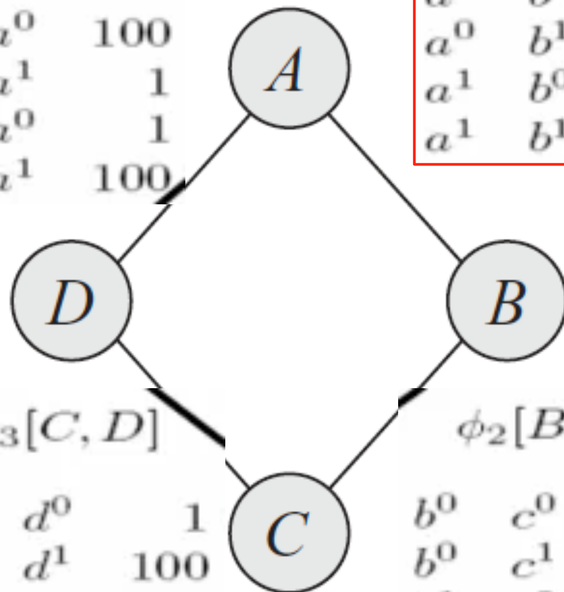
Factors:
Edge
Potentials

a^0 = has misconception
a^1 = no misconception

b^0 = has misconception
b^1 = no misconception

$\phi_4[D, A]$		
d^0	a^0	100
d^0	a^1	1
d^1	a^0	1
d^1	a^1	100

$\phi_1[A, B]$		
a^0	b^0	30
a^0	b^1	5
a^1	b^0	1
a^1	b^1	10



$\phi_3[C, D]$		
c^0	d^0	1
c^0	d^1	100
c^1	d^0	100
c^1	d^1	1

$\phi_2[B, C]$		
b^0	c^0	100
b^0	c^1	1
b^1	c^0	1
b^1	c^1	100

Take negative natural logarithm

$$P(A,B,C,D) \;\alpha \prod_{i=1,2,3,4} \exp\left(-\varepsilon_i(D_i)\right)$$

Factors:
Energy
Functions

$\epsilon_1(A, B)$		
a^0	b^0	-3.4
a^0	b^1	-1.61
a^1	b^0	0
a^1	b^1	-2.3

$\epsilon_2(B, C)$		
b^0	c^0	-4.61
b^0	c^1	0
b^1	c^0	0
b^1	c^1	-4.61

$\epsilon_3(C, D)$		
c^0	d^0	0
c^0	d^1	-4.61
c^1	d^0	-4.61
c^1	d^1	0

$\epsilon_4(D, A)$		
d^0	a^0	-4.61
d^0	a^1	0
d^1	a^0	0
d^1	a^1	-4.61

Log-linear makes potentials apparent

$\epsilon_1(A,B)$			$\epsilon_2(B,C)$			$\epsilon_3(C,D)$			$\epsilon_4(D,A)$		
a^0	b^0	-3.4	b^0	c^0	-4.61	c^0	d^0	0	d^0	a^0	-4.61
a^0	b^1	-1.61	b^0	c^1	0	c^0	d^1	-4.61	d^0	a^1	0
a^1	b^0	0	b^1	c^0	0	c^1	d^0	-4.61	d^1	a^0	0
a^1	b^1	-2.3	b^1	c^1	-4.61	c^1	d^1	0	d^1	a^1	-4.61

$\varepsilon_2(B,C)$ and $\varepsilon_4(D,A)$ take on values that are constant (-4.61) multiples of 1 and 0 for agree/disagree

- Such structure is captured by general framework of features
  - Defined next

Features in a Markov Network

- If $\boldsymbol{D}$ is a subset of variables, feature $f(\boldsymbol{D})$ is a function from $\boldsymbol{D}$ to R (a real value)
- Feature is a factor *without* a non-negativity requirement
- Given a set of k features $\{ f_1(D_1), .. f_k(D_k) \}$

$$\boxed{P(X_1,..,X_n) = \frac{1}{Z}\exp\left[-\sum_{i=1}^{k} w_i f_i(D_i)\right]}$$

  – where $w_i f_i(D_i)$ is entry in energy function table, since

$$P(X_1,..,X_n) = \frac{1}{Z}\exp\left[-\sum_{i=1}^{m} \varepsilon_i(D_i)\right]$$

    - Can have several functions over same scope, $k \neq m$
      – So can represent a standard set of table potentials

Example of Feature

- Pairwise Markov Network $\overset{\frown}{A{-}B{-}C}$
  - Variables are binary
  - Three clusters: $C_1 = \{A,B\},\ C_2=\{B,C\},\ C_3=\{C,A\}$
  - Log-linear model with features
    - $f_{00}(x,y)=1$ if $x=0,\ y=0$; 0 otherwise for x,y instance of C_i
    - $f_{11}(x,y)=1$ if $x=1,\ y=1$ and 0 otherwise
  - Three data instances (A,B,C): $(0,0.0),(0,1,0),(1,0,0)$
    - Unnormalized Feature counts are

$$E_{\tilde{P}}\left[f_{00}\right]=(3+1+1)/3=5/3$$

$$E_{\tilde{P}}\left[f_{11}\right]=(0+0+0)/3=0$$

Definition of Log-linear model with features

- A distribution P is a log-linear model over H if
  - A set of k features $F=\{f_1(D_1),..f_k(D_k)\}$ where each D_i is a complete subgraph and a set of weights w_i
- Such that

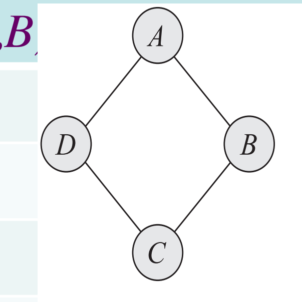
$$\boxed{P(X_1,..,X_n)=\frac{1}{Z}\exp\left[-\sum_{i=1}^{k} w_i f_i(D_i)\right]}$$

  - Note that k is the no of features, not no of subgraphs

Example of binary features

$$\boxed{P(X_1,..X_n;\theta)=\frac{1}{Z(\theta)}\exp\left\{\sum_{i=1}^{k}\theta_i f_i(D_i)\right\}}$$

A	B	$\phi_1(A,B)$
a^0	b^0	$\phi^{a0,b0}$
a^0	b^1	$\phi^{a0,b1}$
a^1	b^0	$\phi^{a1,b0}$
a^1	b^1	$\phi^{a1,b1}$



$Val(A)=\{a^0,a^1\}$ $Val(B)=\{b^0,b^1\}$

$\phi_1(A,B)$ *is* defined for four features $f_{a0,b0}$, $f_{a0,b1}$, $f_{a1,b0}$, and f_{a1b1}

$f_{a0,b0}=1$ *if* $a=a^0, b=b^0$
0 *otherwise, etc.*

- Diamond Network
- With all four variables binary-valued
- Features corresponding to this network are sixteen indicator functions
  - Four for each assignment of variables to four pairwise clusters

$$f_{a^0b^0}(a,b)=I\{a=a^0\}I\{b=b^0\}$$

  - With this representation

$$\boxed{\theta_{a^0b^0}=\ln\phi_1\left(a^0,b^0\right)}$$

Compaction using Features

- Consider $D=\{A_1,A_2\}$ each have l values $a^1,..a^l$
  - As a full factor, clique potential would need l^2 values

ϕ	*Potential*
a^0, b^0	
a^l, b^l	

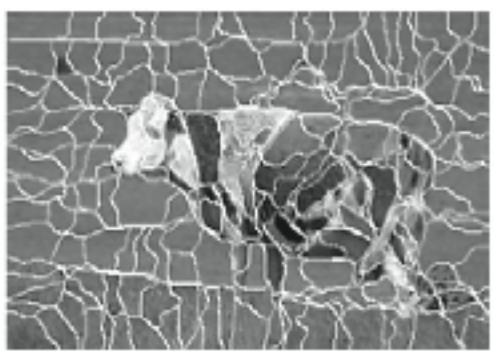
- If we prefer situations in which $A_1=A_2$ but no preference for others, energy function is

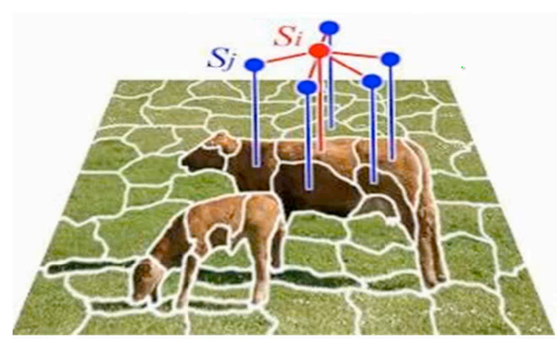
$$\varepsilon(A_1,A_2) = 3 \text{ if } A_1=A_2$$
$$= 0 \text{ otherwise}$$

  – We can encode it as a feature $f(A_1,A_2)$ is an *indicator function* for the event $A_1=A_2$
    - Energy ε is 3 times this feature

Indicator Feature

- A type of feature of particular interest
- Takes on value 1 for some values $\boldsymbol{y} \in Val(\boldsymbol{D})$ and 0 for others
- Example:
  - $\boldsymbol{D} = \{A_1, A_2\}$: each variable has l values $a^1, .. a^l$
  - Function $\phi(A_1, A_2)$ is an indicator function for the event $A_1 = A_2$
    - E.g., two super-pixels have the same greyscale





Neural network and Markov Network

Classification Problem: Features $\boldsymbol{x} = \{x_1, .. x_d\}$ and two-class label y

Neuron(Logistic Regression) is same as a Conditional MN with a single query variable:

feature parameters w_i

Conditional Probability:

Unnormalized $\tilde{P}(y=1 \mid \mathrm{x}) = \exp\left\{w_0 + \sum_{i=1}^{d} w_i x_i\right\}$ $\tilde{P}(y=0 \mid \mathrm{x}) = \exp\{0\} = 1$

Normalized $P(y=1 \mid \mathrm{x}) = sigmoid\left\{w_0 + \sum_{i=1}^{d} w_i x_i\right\}$ where $sigmoid(z) = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}}$

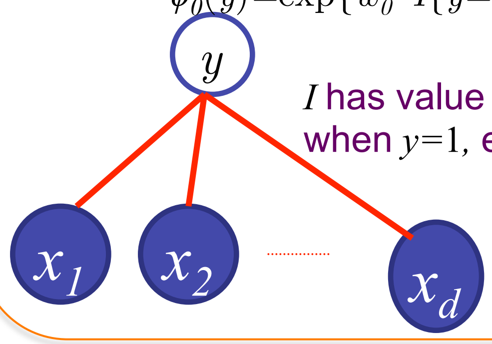
Z has term 1 because $\tilde{P}(y=0 \mid \mathrm{x})=1$

Factors (log-linear w. features):

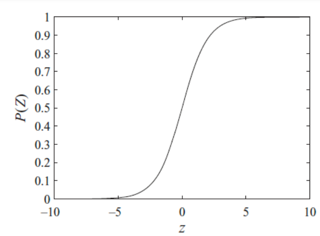
$D_i = \{x_i, y\}$ $f_i(D_i) = x_i\, I(y)$

$\phi_i(x_i, y) = \exp\{w_i x_i\ I\{y=1\}\}$,

$\phi_0(y) = \exp\{w_0\ I\{y=1\}\}$



I has value 1 when $y=1$, else 0



sigmoid

Learning: Jointly optimize d parameters w_i
High dimensional estimation
but correlations accounted for
Can use much richer features:
Edges, image patches sharing same pixels

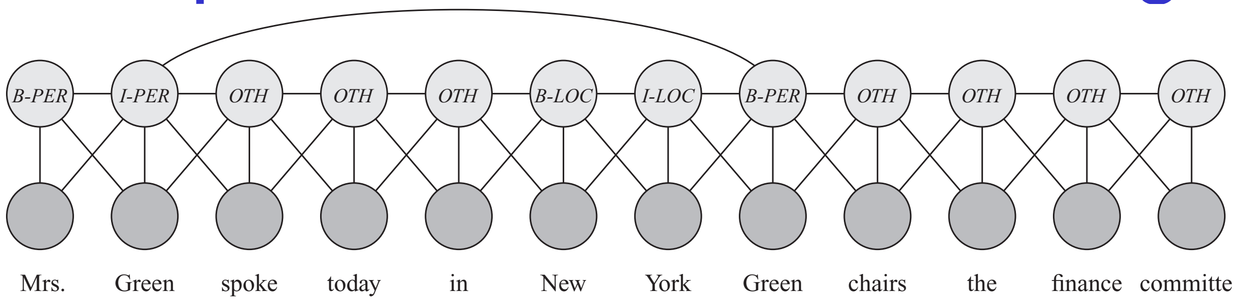
C-class

$$p(y_c \mid \mathrm{x}) = \frac{\exp(\mathrm{w}_c^T \mathrm{x})}{\sum_j^C \exp(\mathrm{w}_j^T \mathrm{x})}$$

$C \times d$ parameters

Use of Features in Text Analysis

- Compact for variables with large domains



$\boldsymbol{Y}=$ target variables

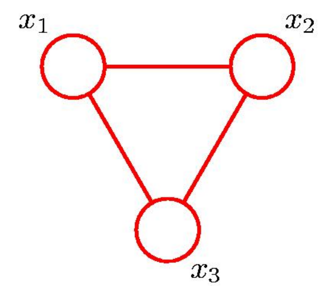
$\boldsymbol{X}=$ known variables

  - $\boldsymbol{X}$ are words of text, $\boldsymbol{Y}$ are named entities
    - *B-PER*=Begin Person, *I-PER*=within person, *OTH*=Not entity
- Factors for word t: $\Phi^1{}_t(Y_t, Y_{t+1})$, $\Phi^2{}_t(Y_t, X_1, ..X_T)$
- Features (hundreds of thousands):
    - Word itself (capitalised, in list of common names)
    - Aggregate features of sequence (>2 sports related)
    - Y_t dependent on several words in a window of t
    - Skip chain CRF: connections between adjacent words & multiple occurences of same word

Examples of Feature Parameterization of MNs

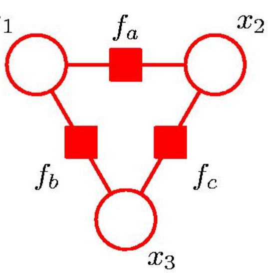
- Text Analysis
- Ising Model
- Boltzmann Model
- Metric MRFs

Summary of three MN parameterizations (each finer than previous)



1. Markov network

- Product of potentials on cliques
- Good for discussing independence queries



2. Factor Graphs

- Product of factors describes Gibbs distribution
- Useful for inference

$$P_\Phi(X_1,..X_n) = \frac{1}{Z}\tilde{P}(X_1,..X_n) \text{ where } \tilde{P}(X_1,..X_n) = \prod_{i=1}^{m}\phi_i(D_i)$$

is an unnomalized measure and $Z = \sum_{X_1,..X_n} \tilde{P}(X_1,..X_n)$

3. Features

- Product of features
- Can describe all entries in each factor
- For both hand-coded models and for learning

$$P(X_1,..,X_n) = \frac{1}{Z}\exp\left[-\sum_{i=1}^{k} w_i f_i(D_i)\right]$$