Readings:
K&F: 9.1, 9.2, 9.3, 9.4
K&F: 5.1, 5.2, 5.3, 5.4, 5.5, 5.6

Clique Trees 3

Let's get BP right

Undirected Graphical Models

Here the couples get to swing!

Graphical Models – 10708

Carlos Guestrin

Carnegie Mellon University

October 25th, 2006

Factor division

- Let **X** and **Y** be disjoint set of variables
- Consider two factors: $\phi_1(\mathbf{X},\mathbf{Y})$ and $\phi_2(\mathbf{Y})$
- Factor $\psi=\phi_1/\phi_2$
  - 0/0=0

$\phi_1(A,B)$

a^1	b^1	0.5
a^1	b^2	0.2
a^2	b^1	0
a^2	b^2	0
a^3	b^1	0.3
a^3	b^2	0.45

$\phi_2(A)$

a^1	0.8
a^2	0
a^3	0.6

$\psi(A,B)$

a^1	b^1	0.625
a^1	b^2	0.25
a^2	b^1	0
a^2	b^2	0
a^3	b^1	0.5
a^3	b^2	0.75

$$\psi(X=x,Y=y)=\frac{\phi_1(X=x,Y=y)}{\phi_2(Y=y)}$$

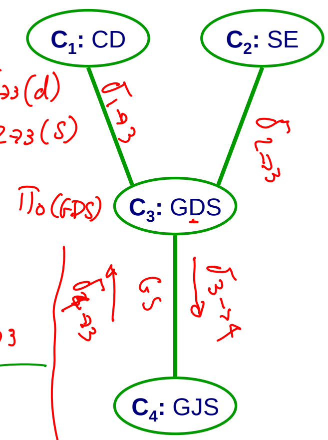
$$\psi(A=a^1,B=b^2)=\frac{0.2}{0.8}=0.25$$

Introducing message passing with division

- Variable elimination (message passing with multiplication)
  - message: $\delta_{3\to4}(G,S) = \sum_d \pi_0(G,d,S)\cdot\delta_{1\to3}(d)\cdot\delta_{2\to3}(S)$
  - belief: $\pi_3(GDS) = \pi_0(GDS)\,\delta_{1\to3}\cdot\delta_{2\to3}\cdot\delta_{4\to3}$
- Message passing with division:
  - message: Estimate of $\pi_3(GDS) = \pi_0\cdot\delta_{1\to3}\cdot\delta_{2\to3}\cdot\delta_{4\to3}$

    $\delta_{3\to4}(GS) = \dfrac{\sum_d \pi_3(GdS)}{\delta_{4\to3}(GS)}$
  - belief update:

$\pi_4(GJS) \leftarrow \pi_4(GJS)\cdot\delta_{3\to4}(GS)$

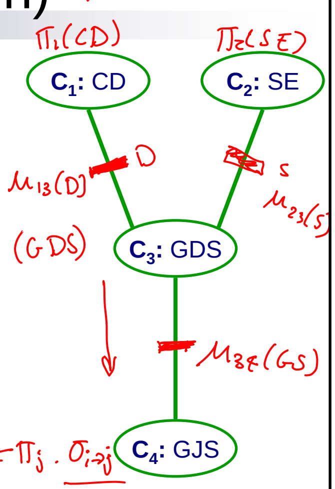


Lauritzen-Spiegelhalter Algorithm (a.k.a. belief propagation)

initialized to $\pi_0(CD)$

- Separator potentials μ_{ij}
  - one per edge (same both directions)
  - holds "last message"
  - initialized to 1
- Message $i\to j$
  - what does i think the separator potential should be?
    - $\sigma_{i\to j}(S_{ij}) = \sum_{C_i/S_{ij}} \pi_i(C_i)$
  - update belief for j:
    - pushing j to what i thinks about separator $\pi_j \leftarrow \pi_j\cdot\dfrac{\sigma_{i\to j}}{\mu_{ij}}$
  - replace separator potential:

    $\mu_{ij} \leftarrow \sigma_{i\to j}$



Clique tree invariant

- **Clique tree potential**: $\pi_T(X) = \dfrac{\prod_i \pi_i(C_i)}{\prod_{ij} \mu_{ij}(S_{ij})}$
  - Product of clique potentials divided by separators potentials

initialization: $\mu_{ij} = 1$, $\pi_i(C_i) = \pi_0(C_i)$

$$\pi_T(X) = \frac{\prod_i \pi_0(C_i)}{\prod_{ij} 1} = \prod_j P(X_j \mid Pa_{X_j}) = P(X)$$

- Clique tree **invariant**:
  - $P(\mathbf{X}) = \pi_T(\mathbf{X})$

$P(X) = \pi_T(X)$ before message

$= \pi'_T(X)$ after message

use π_j, μ_{ij} ?

send message $i \to j$

$\pi'_j \leftarrow \pi_j \cdot \dfrac{\sigma_{i \to j}}{\mu_{ij}}$

$\mu'_{ij} \leftarrow \sigma_{i \to j}$

use $\dfrac{\pi'_j}{\mu'_{ij}}$

$\dfrac{\pi'_j}{\mu'_{ij}} = \dfrac{\pi_j \cdot \frac{\sigma_{i \to j}}{\mu_{ij}}}{\sigma_{i \to j}} = \dfrac{\pi_j}{\mu_{ij}}$

Belief propagation and clique tree invariant

- **Theorem**: Invariant is maintained by BP algorithm!

$\pi_T(X) = P(X)$

at convergence: $\pi_i(C_i) = P(C_i)$

$\mu_{ij}(S_{ij}) = P(S_{ij})$

- BP reparameterizes clique potentials and separator potentials
  - At convergence, potentials and messages are marginal distributions

$$P(X) = \frac{\prod_i P(C_i)}{\prod_{ij} P(S_{ij})}$$

Subtree correctness

- **Informed message** from i to j, if all messages into i (other than from j) are informed
  - Recursive definition (leaves always send informed messages)
- **Informed subtree**:
  - All incoming messages informed
- **Theorem**:
  - Potential of connected informed subtree *T'* is marginal over scope[*T'*]
- **Corollary**:
  - At convergence, clique tree is *calibrated*
    - $\pi_i = P(\text{scope}[\pi_i])$
    - $\mu_{ij} = P(\text{scope}[\mu_{ij}])$

Clique trees versus VE

- Clique tree advantages
  - Multi-query settings
  - Incremental updates
  - Pre-computation makes complexity explicit
- Clique tree disadvantages
  - Space requirements – no factors are "deleted"
  - Slower for single query
  - Local structure in factors may be lost when they are multiplied together into initial clique potential

Clique tree summary

- Solve marginal queries for all variables in only twice the cost of query for one variable
- Cliques correspond to maximal cliques in induced graph
- Two message passing approaches
  - VE (the one that multiplies messages)
  - BP (the one that divides by old message)
- Clique tree invariant
  - Clique tree potential is always the same
  - We are only reparameterizing clique potentials
- Constructing clique tree for a BN
  - from elimination order
  - from triangulated (chordal) graph
- Running time (only) exponential in size of largest clique
  - Solve **exactly** problems with thousands (or millions, or more) of variables, and cliques with tens of nodes (or less)

Announcements

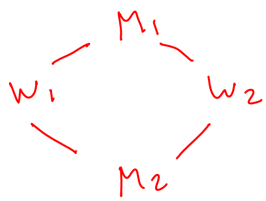
- Recitation tomorrow, don't miss it!!!
  - Khalid on Undirected Models

Swinging Couples revisited

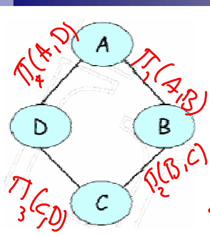
- This is no perfect map in BNs
- But, an undirected model will be a perfect map

$W_1 \perp W_2 \mid M_1, M_2$

$M_1 \perp M_2 \mid W_1, W_2$



Potentials (or Factors) in Swinging Couples



π^0 π^0 π^0 π^0

$\pi_1[A,B]$			$\pi_2[B,C]$			$\pi_3[C,D]$			$\pi_4[D,A]$		
a^0	b^0	30	b^0	c^0	100	c^0	d^0	1	d^0	a^0	100
a^0	b^1	5	b^0	c^1	1	c^0	d^1	100	d^0	a^1	1
a^1	b^0	1	b^1	c^0	1	c^1	d^0	100	d^1	a^0	1
a^1	b^1	10	b^1	c^1	100	c^1	d^1	1	d^1	a^1	100

$\pi_1(A,B) \geq 0 \quad \forall A,B$

$$P(A,B,C,D) = \frac{1}{Z} \pi_1(A,B) \cdot \pi_2(B,C) \cdot \pi_3(C,D) \cdot \pi_4(D,A)$$

normalization

$$Z = \sum_{a,b,c,d} \pi_1(a,b) \cdot \pi_2(b,c) \cdot \pi_3(c,d) \cdot \pi_4(d,a)$$

Computing probabilities in Markov networks v. BNs

- In a BN, can compute prob. of an instantiation by multiplying CPTs

$P(A=a, B=b, C=c, D=d)$
$= P(A=a) \cdot P(B=b|A=a) \cdot P(C=c|A=a) \cdot P(D=d|B=b, C=c)$

Can't do this directly in a MN, you need to know Z

- In an Markov networks, can only compute ratio of probabilities directly

$\pi_1[A,B]$			$\pi_2[B,C]$			$\pi_3[C,D]$			$\pi_4[D,A]$		
a^0	b^0	30	b^0	c^0	100	c^0	d^0	1	d^0	a^0	100
a^0	b^1	5	b^0	c^1	1	c^0	d^1	100	d^0	a^1	1
a^1	b^0	1	b^1	c^0	1	c^1	d^0	100	d^1	a^0	1
a^1	b^1	10	b^1	c^1	100	c^1	d^1	1	d^1	a^1	100

$$\frac{P(A=t, B=t, C=t, D=t)}{P(A=f, B=f, C=f, D=f)} = \frac{\cancel{\frac{1}{Z}}\ \underset{10}{\pi_1(t,t)}\ \underset{100}{\pi_2(t,t)}\ \underset{1}{\pi_3(tt)}\ \underset{100}{\pi_4(tt)}}{\cancel{\frac{1}{Z}}\ \underset{30}{\pi_1(ff)} \cdot \underset{100}{\pi_2(ff)} \cdot \underset{1}{\pi_3(ff)}\ \underset{100}{\pi_4(ff)}}$$

$= 1/3$

Normalization for computing probabilities

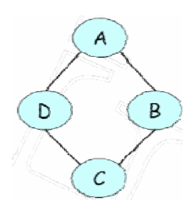
- To compute actual probabilities, must compute normalization constant (also called partition function)

$P(ABCD) = \frac{1}{Z}\ \pi_1(AB)\ \pi_2(BC)\ \pi_3(CD)\ \pi_4(DA)$

Assignment				*Unnormalized*	*Normalized*
a^0	b^0	c^0	d^0	300000	0.04
a^0	b^0	c^0	d^1	300000	0.04
a^0	b^0	c^1	d^0	300000	0.04
a^0	b^0	c^1	d^1	30	$4.1 \cdot 10^{-6}$
a^0	b^1	c^0	d^0	500	$6.9 \cdot 10^{-5}$
a^0	b^1	c^0	d^1	500	$6.9 \cdot 10^{-5}$
a^0	b^1	c^1	d^0	5000000	0.69
a^0	b^1	c^1	d^1	500	$6.9 \cdot 10^{-5}$
a^1	b^0	c^0	d^0	100	$1.4 \cdot 10^{-5}$
a^1	b^0	c^0	d^1	1000000	0.14
a^1	b^0	c^1	d^0	100	$1.4 \cdot 10^{-5}$
a^1	b^0	c^1	d^1	100	$1.4 \cdot 10^{-5}$
a^1	b^1	c^0	d^0	10	$1.4 \cdot 10^{-6}$
a^1	b^1	c^0	d^1	100000	0.014
a^1	b^1	c^1	d^0	100000	0.014
a^1	b^1	c^1	d^1	100000	0.014

- Computing partition function is hard! → Must sum over all possible assignments

$Z = \sum_{abcd} \pi_1(ab)\ \pi_2(bc) \cdot \pi_3(cd)\ \pi_4(da)$

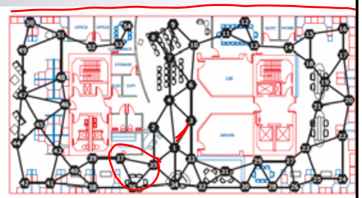


Factorization in Markov networks

- Given an undirected graph *H* over variables $\mathbf{X}=\{X_1,\ldots,X_n\}$
- A distribution *P* **factorizes** over *H* if ∃
  - subsets of variables $\mathbf{D}_1 \subseteq \mathbf{X}, \ldots, \mathbf{D}_m \subseteq \mathbf{X}$, such that the $\mathbf{D}_i$ are *fully connected* in *H*
  - *non-negative potentials* (or factors) $\pi_1(\mathbf{D}_1), \ldots, \pi_m(\mathbf{D}_m)$
    - also known as clique potentials
  - such that

$$P(X) = \frac{1}{Z} \prod_{i=1}^{m} \pi_i(D_i)$$

- Also called Markov random field *H*, or Gibbs distribution over *H*



35, 36, 37

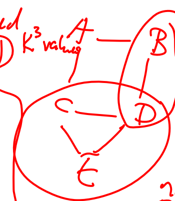
allowed

Stronger $\pi(CDE)$ K^3 values

or

Weaker $\pi(CD)$ $\pi(DE)$ $\pi(CE)$

$3K^2$ values



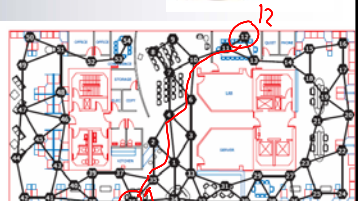
not allowed

no factor AB

no factor BCD

Global Markov assumption in Markov networks

- A path $X_1 - \ldots - X_k$ is **active** when set of variables **Z** are observed if none of $X_i \in \{X_1,\ldots,X_k\}$ are observed (are part of **Z**)
- Variables **X** are **separated** from **Y** given **Z** in graph *H*, $sep_H(\mathbf{X};\mathbf{Y}|\mathbf{Z})$, if there is no active path between any $X \in \mathbf{X}$ and any $Y \in \mathbf{Y}$ given **Z**
- The **global Markov assumption** for a Markov network *H* is $I(H) = \{X, Y, Z \mid sep_H(X;Y|Z)\}$



12

36

active if I don't observe 35, 1, 4...

The BN Representation Theorem

$I(G) \subseteq I(P)$ I-map

If conditional independencies in BN are subset of conditional independencies in *P*

Obtain

Joint probability distribution:

$$P(X_1,\ldots,X_n) = \prod_{i=1}^{n} P\left(X_i \mid \mathbf{Pa}_{X_i}\right)$$

Important because:
Independencies are sufficient to obtain BN structure *G*

If joint probability distribution:

$$P(X_1,\ldots,X_n) = \prod_{i=1}^{n} P\left(X_i \mid \mathbf{Pa}_{X_i}\right)$$

Obtain

$I(G) \subseteq I(P)$

Then conditional independencies in BN are subset of conditional independencies in *P*

Important because:
Read independencies of *P* from BN structure *G*

Markov networks representation Theorem 1

If joint probability distribution *P*:

$$P(X_1,\ldots,X_n) = \frac{1}{Z}\prod_{i=1}^{m} \pi_i\left(\mathbf{D}_i\right)$$

Then

H* is an I-map for *P

$I(H) \subseteq I(P)$

P factorizes according to H

- If you can write distribution as a normalized product of factors ⇒ Can read independencies from graph

What about the other direction for Markov networks ?

If *H* is an I-map for *P* ~~**Then**~~ **joint probability distribution *P*:** $P(X_1,\ldots,X_n) = \frac{1}{Z}\prod_{i=1}^{m}\pi_i(\mathbf{D}_i)$

$I(H) \subseteq I(P)$

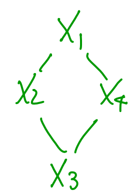
not true!!

- Counter-example: $X_1,\ldots,X_4$ are binary, and only eight assignments have positive probability: P factorizes according to H

  (0,0,0,0) 1/8 (1,0,0,0) 1/8 (1,1,0,0) (1,1,1,0)
  (0,0,0,1) (0,0,1,1) (0,1,1,1) (1,1,1,1)

  $P(0,1,0,1) = 0$

- For example, $X_1 \perp X_3 | X_2, X_4$: $P(X_1=0 | X_2=0, X_4=0) = 1/2$ indep of value of X_3



- But distribution doesn't factorize!!! Homework !!

Markov networks representation Theorem 2 (Hammersley-Clifford Theorem)

If *H* is an I-map for *P* and *P* is a positive distribution **Then** **joint probability distribution *P*:** $P(X_1,\ldots,X_n) = \frac{1}{Z}\prod_{i=1}^{m}\pi_i(\mathbf{D}_i)$

$\forall x \quad P(x) > 0$

- Positive distribution and independencies ⇒ *P* factorizes over graph

Representation Theorem for Markov Networks

If joint probability distribution P:

$$P(X_1, \ldots, X_n) = \frac{1}{Z} \prod_{i=1}^{m} \pi_i(\mathbf{D}_i)$$

Then

H is an I-map for *P*

If *H* is an I-map for *P* and *P* is a positive distribution

Then

joint probability distribution *P*:

$$P(X_1, \ldots, X_n) = \frac{1}{Z} \prod_{i=1}^{m} \pi_i(\mathbf{D}_i)$$

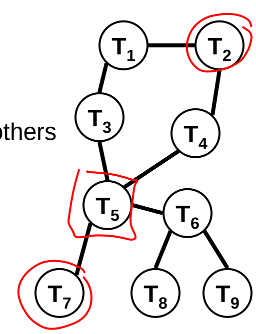
Completeness of separation in Markov networks

- **Theorem: Completeness of separation**
  - For "almost all" distributions that *P* factorize over Markov network *H*, we have that I(*H*) = I(*P*)
  - you know that $I(H) \subseteq I(P)$
  - *"almost all" distributions*: except for a set of measure zero of parameterizations of the Potentials (assuming no finite set of parameterizations has positive measure)
- **Analogous to BNs**

What are the “local” independence assumptions for a Markov network?

- In a BN *G*:
  - local Markov assumption: variable independent of non-descendants given parents $I_\ell(G)$
  - d-separation defines global independence $I(G)$
  - Soundness: For all distributions: $P \models I_\ell(G) \Leftrightarrow P \models I(G)$
- In a Markov net *H*:
  - **Separation** defines global independencies
  - What are the notions of local independencies?

Local independence assumptions for a Markov network



- **Separation** defines global independencies $I(H)$
- **Pairwise Markov Independence**: $I_{pw}(H)$
  - Pairs of non-adjacent variables A,B are independent given all others

$$A \perp B \mid \mathcal{X} - \{A,B\}$$

- **Markov Blanket**: $I_{MB}(H)$
  - Variable A independent of rest given its neighbors $N(A)$

$$A \perp \mathcal{X} - \{N(A), A\} \mid N(A)$$

$T_7 \perp T_2 \mid T_1, T_{3-6}, T_8, T_9$

$T_7 \perp \{T_1, T_{2-6}, T_{8-9}\} \mid T_5$

Equivalence of independencies in Markov networks

- **Soundness Theorem**: For all positive distributions *P*, the following three statements are equivalent:
  - *P* entails the global Markov assumptions

    $P \models I(H)$
  - *P* entails the pairwise Markov assumptions

    $P \models I_{PW}(H)$
  - *P* entails the local Markov assumptions (Markov blanket)

    $P \models I_{MB}(H)$

Minimal I-maps and Markov Networks

- A fully connected graph is an I-map
- Remember minimal I-maps?
  - A "simplest" I-map → Deleting an edge makes it no longer an I-map
- In a BN, there is no unique minimal I-map
- Theorem: **In a Markov network, minimal I-map is unique!!**
- Many ways to find minimal I-map, e.g.,
  - Take pairwise Markov assumption:
  - If *P* doesn't entail it, add edge:

How about a perfect map?

- Remember perfect maps?
  - independencies in the graph are exactly the same as those in *P*
- For BNs, doesn't always exist
  - counter example: Swinging Couples
- How about for Markov networks?

Unifying properties of BNs and MNs

- BNs:
  - give you: V-structures, CPTs are conditional probabilities, can directly compute probability of full instantiation
  - but: require acyclicity, and thus no perfect map for swinging couples
- MNs:
  - give you: cycles, and perfect maps for swinging couples
  - but: don't have V-structures, cannot interpret potentials as probabilities, requires partition function
- Remember PDAGS???
  - skeleton + immoralities
  - provides a (somewhat) unified representation
  - see book for details

What you need to know so far about Markov networks

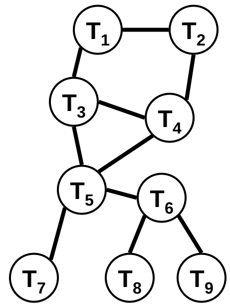
- Markov network representation:
  - undirected graph
  - potentials over cliques (or sub-cliques)
  - normalize to obtain probabilities
  - need partition function
- Representation Theorem for Markov networks
  - if P factorizes, then it's an I-map
  - if P is an I-map, only factorizes for positive distributions
- Independence in Markov nets:
  - active paths and separation
  - pairwise Markov and Markov blanket assumptions
  - equivalence for positive distributions
- Minimal I-maps in MNs are unique
- Perfect maps don't always exist

Some common Markov networks and generalizations

- Pairwise Markov networks
- A very simple application in computer vision
- Logarithmic representation
- Log-linear models
- Factor graphs

Pairwise Markov Networks

- All factors are over single variables or pairs of variables:
  - Node potentials
  - Edge potentials
- Factorization:



- Note that there may be bigger cliques in the graph, but only consider pairwise potentials

A very simple vision application

- Image segmentation: separate foreground from background
- Graph structure:
  - pairwise Markov net
  - grid with one node per pixel
- Node potential:
  - "background color" v. "foreground color"
- Edge potential:
  - neighbors like to be of the same class

Logarithmic representation

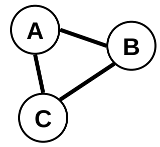
- Standard model: $P(X_1,\ldots,X_n) = \frac{1}{Z}\prod_{i=1}^{m}\pi_i(\mathbf{D}_i)$
- Log representation of potential (assuming positive potential):
  - also called the energy function
- Log representation of Markov net:

Log-linear Markov network (most common representation)

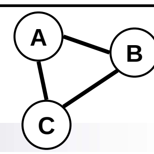
- **Feature** is some function $\phi[\mathbf{D}]$ for some subset of variables **D**
  - e.g., indicator function
- **Log-linear model** over a Markov network *H*:
  - a set of features $\phi_1[\mathbf{D}_1],\ldots,\phi_k[\mathbf{D}_k]$
    - each $\mathbf{D}_i$ is a subset of a clique in *H*
    - two ϕ's can be over the same variables
  - a set of weights $w_1,\ldots,w_k$
    - usually learned from data
  - $P(X_1,\ldots,X_n) = \frac{1}{Z}\exp\left[\sum_{i=1}^{k} w_i\phi_i(\mathbf{D}_i)\right]$

Structure in cliques

- Possible potentials for this graph:



Factor graphs



- Very useful for approximate inference
  - Make factor dependency explicit
- Bipartite graph:
  - variable nodes (ovals) for $X_1,\ldots,X_n$
  - factor nodes (squares) for $\phi_1,\ldots,\phi_m$
  - edge $X_i - \phi_j$ if $X_i \in \text{Scope}[\phi_j]$

Summary of types of Markov nets

- Pairwise Markov networks
  - very common
  - potentials over nodes and edges
- Log-linear models
  - log representation of potentials
  - linear coefficients learned from data
  - most common for learning MNs
- Factor graphs
  - explicit representation of factors
    - you know exactly what factors you have
  - very useful for approximate inference