Readings:
K&F: 4.1, 4.2, 4.3, 4.4, 4.5

Undirected Graphical Models (finishing off)

Graphical Models – 10708
Carlos Guestrin
Carnegie Mellon University
November 3rd, 2008

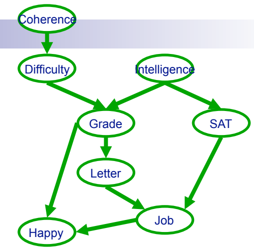
What you learned about so far

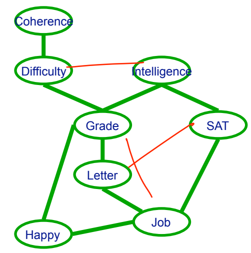
- Bayes nets
- Junction trees
- (General) Markov networks
- Pairwise Markov networks
- Factor graphs

- How do we transform between them?
- More formally:
  - I give you an graph in one representation, find an **I-map** in the other

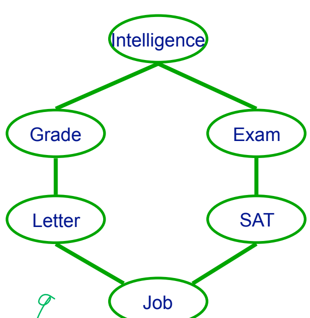
BNs → MNs: Moralization

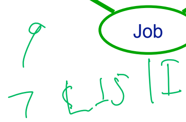
- **Theorem**: Given a BN *G* the Markov net *H* formed by moralizing *G* is the *minimal I-map* for I(*G*)
- **Intuition**:
  - in a Markov net, each factor must correspond to a subset of a clique
  - the factors in BNs are the CPTs
  - CPTs are factors over a node and its parents
  - thus node and its parents must form a clique
- **Effect**:
  - **some** independencies that could be read from the BN graph become hidden

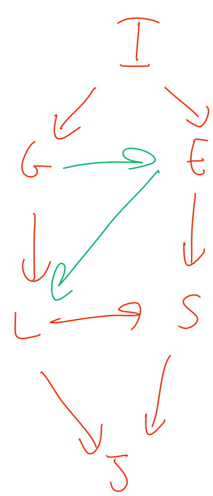




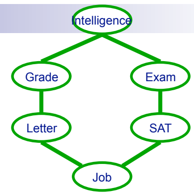
From Markov nets to Bayes nets

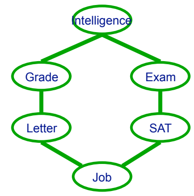






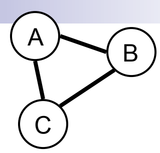
MNs → BNs: Triangulation





- **Theorem**: Given a MN *H,* let *G* be the Bayes net that is a *minimal I-map* for I(*H)* then *G* must be **chordal**
- **Intuition**:
  - v-structures in BN introduce immoralities
  - these immoralities were not present in a Markov net
  - the triangulation eliminates immoralities
- **Effect**:
  - **many** independencies that could be read from the MN graph become hidden

Markov nets v. Pairwise MNs



- Every Markov network can be transformed into a Pairwise Markov net
  - introduce extra “variable” for each factor over three or more variables
  - domain size of extra variable is exponential in number of vars in factor
- **Effect**:
  - any local structure in factor is lost
  - a chordal MN doesn’t look chordal anymore

Overview of types of graphical models and transformations between them

Readings:
K&F: 10.1, 10.5

Mean Field and Variational Methods

First approximate inference

Graphical Models – 10708
Carlos Guestrin
Carnegie Mellon University

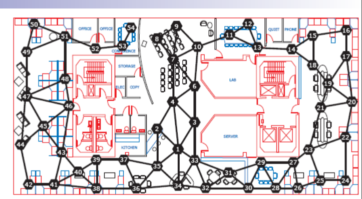
November 3rd, 2008

Approximate inference overview

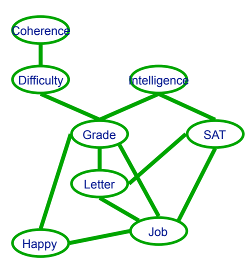
- So far: VE & junction trees
  - exact inference
  - exponential in tree-width
- There are many many many many approximate inference algorithms for PGMs
- We will focus on three representative ones:
  - sampling
  - variational inference
  - loopy belief propagation and generalized belief propagation

Approximating the posterior v. approximating the prior

- Prior model represents entire world
  - world is complicated
  - thus prior model can be very complicated



- Posterior: after making observations
  - sometimes can become much more sure about the way things are
  - sometimes can be approximated by a simple model
- First approach to approximate inference: **find *simple* model that is "close" to posterior**
- Fundamental problems:
  - **what is close?**
  - **posterior is intractable result of inference, how can we approximate what we don't have?**



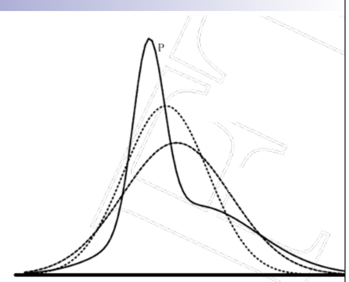
KL divergence: Distance between distributions

- Given two distributions *p* and *q* KL divergence:

- D(p||q) = 0 iff p=q
- Not symmetric – p determines where difference is important
  - p(x)=0 and q(x)≠0
  - p(x)≠0 and q(x)=0

Find simple approximate distribution

- Suppose *p* is intractable posterior
- Want to find simple *q* that approximates *p*
- KL divergence not symmetric
- D(p||q)
  - true distribution p defines support of diff.
  - the "correct" direction
  - will be intractable to compute
- D(q||p)
  - approximate distribution defines support
  - tends to give overconfident results
  - will be tractable



Back to graphical models

- Inference in a graphical model:
  - P(**x**) =
  - want to compute $P(X_i|\mathbf{e})$
  - our *p*:
- What is the simplest *q*?
  - every variable is independent:
  - mean field approximation
  - can compute any prob. very efficiently

D(p||q) for mean field – KL the right way

- p:
- q:
- D(p||q)=

D(q||p) for mean field – KL the reverse direction

- p:
- q:
- D(q||p)=

D(q||p) for mean field – KL the reverse direction: Entropy term

- p:
- q:

$$D(q||p) = \sum_x q(x) \log q(x) - \sum_x q(x) \log p(x)$$

D(q||p) for mean field – KL the reverse direction: cross-entropy term

- p:
- q:

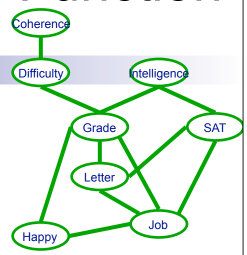
$$D(q||p) = \sum_x q(x) \log q(x) - \sum_x q(x) \log p(x)$$

What you need to know so far

- Goal:
  - Find an efficient distribution that is close to posterior
- Distance:
  - measure distance in terms of KL divergence
- Asymmetry of KL:
  - D(p||q) ≠ D(q||p)
- Computing right KL is intractable, so we use the reverse KL

Reverse KL & The Partition Function

Back to the general case



- Consider again the defn. of D(q||p):
  - p is Markov net P_F

- **Theorem**: $\ln Z = F[P_{\mathcal{F}}, Q] + D(Q||P_{\mathcal{F}})$

- where energy functional:

$$F[P_{\mathcal{F}}, Q] = \sum_{\phi \in \mathcal{F}} E_Q[\ln \phi] + H_Q(\mathcal{X})$$

Understanding Reverse KL, Energy Function & The Partition Function

$\ln Z = F[P_{\mathcal{F}}, Q] + D(Q||P_{\mathcal{F}})$ $F[P_{\mathcal{F}}, Q] = \sum_{\phi \in \mathcal{F}} E_Q[\ln \phi] + H_Q(\mathcal{X})$

- Maximizing Energy Functional ⇔ Minimizing Reverse KL

- **Theorem**: Energy Function is lower bound on partition function
  - Maximizing energy functional corresponds to search for tight lower bound on partition function

Structured Variational Approximate Inference

$\ln Z = F[P_{\mathcal{F}}, Q] + D(Q||P_{\mathcal{F}})$

$F[P_{\mathcal{F}}, Q] = \sum_{\phi \in \mathcal{F}} E_Q[\ln \phi] + H_Q(\mathcal{X})$

- Pick a family of distributions Q that allow for exact inference
  - e.g., fully factorized (mean field)
- Find Q∈Q that maximizes $F[P_{\mathcal{F}}, Q]$
- For mean field

Optimization for mean field

$$\max_Q F[P_{\mathcal{F}}, Q] = \max_Q \sum_{\phi \in \mathcal{F}} E_Q[\ln \phi] + \sum_j H_{Q_j}(X_j)$$

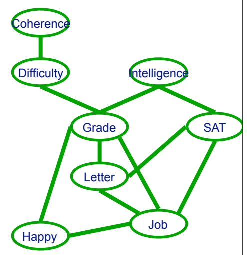
$$\forall i, \sum_{x_i} Q_i(x_i) = 1$$

- Constrained optimization, solved via Lagrangian multiplier
  - ∃ λ, such that optimization equivalent to:
  - Take derivative, set to zero
- **Theorem**: Q is a stationary point of mean field approximation iff for each *i*:

$$Q_i(x_i) = \frac{1}{Z_i} \exp \left\{ \sum_{\phi \in \mathcal{F}} E_Q[\ln \phi \mid x_i] \right\}$$

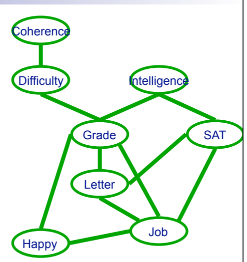
Understanding fixed point equation

$$Q_i(x_i) = \frac{1}{Z_i} \exp\left\{ \sum_{\phi \in \mathcal{F}} E_Q[\ln \phi \mid x_i] \right\}$$



Simplifying fixed point equation

$$Q_i(x_i) = \frac{1}{Z_i} \exp\left\{ \sum_{\phi \in \mathcal{F}} E_Q[\ln \phi \mid x_i] \right\}$$



Q_i only needs to consider factors that intersect X_i

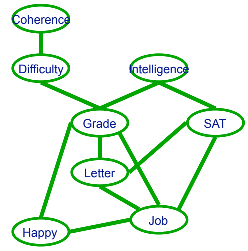
- **Theorem**: The fixed point:

$$Q_i(x_i) = \frac{1}{Z_i} \exp\left\{ \sum_{\phi \in \mathcal{F}} E_Q[\ln \phi \mid x_i] \right\}$$

is equivalent to:

$$Q_i(x_i) = \frac{1}{Z_i} \exp\left\{ \sum_{\phi_j : X_i \in \text{Scope}[\phi_j]} E_Q[\ln \phi_j(\mathbf{U}_j, x_i)] \right\}$$

  - where the Scope[ϕ_j] = $\mathbf{U}_j \cup \{X_i\}$



There are many stationary points!

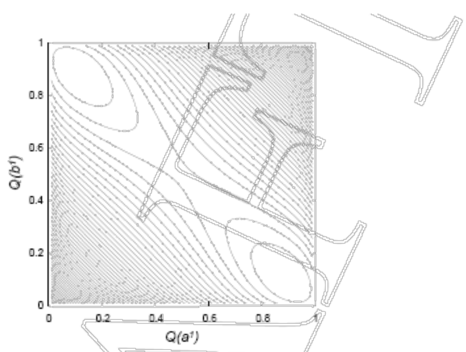
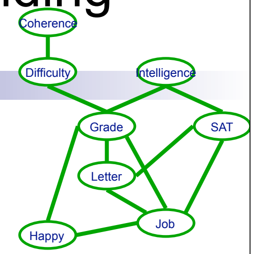


Figure 11.18 An example of a multi-modal mean field energy functional landscape. In this network, $P(a,b) = 0.25 - \epsilon$ if $a \neq b$ and ϵ if $a = b$. The axes correspond to the mean field marginal for A and B and the contours show equi-values of the energy functional.

Very simple approach for finding one stationary point

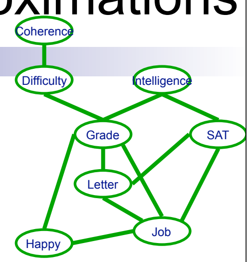


- Initialize Q (e.g., randomly or smartly)
- Set all vars to unprocessed
- Pick unprocessed var X_i
  - update Q_i:

$$Q_i(x_i) = \frac{1}{Z_i} \exp \left\{ \sum_{\phi_j : X_i \in \text{Scope}[\phi_j]} E_Q[\ln \phi_j(\mathbf{U}_j, x_i)] \right\}$$

  - set var i as processed
  - if Q_i changed
    - set neighbors of X_i to unprocessed
- Guaranteed to converge

More general structured approximations



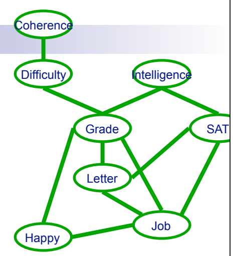
- Mean field very naïve approximation
- Consider more general form for Q
  - assumption: exact inference doable over Q
- **Theorem**: stationary point of energy functional:

$$\psi_j(\mathbf{c_j}) \propto \exp \left\{ \sum_{\phi \in \mathcal{F}} E_Q[\ln \phi \mid \mathbf{c_j}] - \sum_{\psi \in \mathcal{Q} \setminus \{\psi_j\}} E_Q[\ln \psi \mid \mathbf{c_j}] \right\}$$

- Very similar update rule

Computing update rule for general case

$$\psi_j(\mathbf{c_j}) \propto \exp\left\{\sum_{\phi\in\mathcal{F}} E_Q[\ln\phi \mid \mathbf{c_j}] - \sum_{\psi\in\mathcal{Q}\setminus\{\psi_j\}} E_Q[\ln\psi \mid \mathbf{c_j}]\right\}$$



- Consider one ϕ:

Structured Variational update requires inference

$$\psi_j(\mathbf{c_j}) \propto \exp\left\{\sum_{\phi\in\mathcal{F}} E_Q[\ln\phi \mid \mathbf{c_j}] - \sum_{\psi\in\mathcal{Q}\setminus\{\psi_j\}} E_Q[\ln\psi \mid \mathbf{c_j}]\right\}$$

- Compute marginals wrt Q of cliques in original graph and cliques in new graph, for all cliques
- What is a good way of computing all these marginals?
- Potential updates:
  - sequential: compute marginals, update ψ_j, recompute marginals
  - parallel: compute marginals, update all ψ's, recompute marginals

What you need to know about variational methods

- Structured Variational method:
  - select a form for approximate distribution
  - minimize reverse KL
- Equivalent to maximizing energy functional
  - searching for a tight lower bound on the partition function
- Many possible models for *Q*:
  - independent (mean field)
  - structured as a Markov net
  - cluster variational
- Several subtleties outlined in the book