Readings:
K&F: 11.1, 11.5

Mean Field and Variational Methods

First approximate inference

Graphical Models – 10708
Carlos Guestrin
Carnegie Mellon University

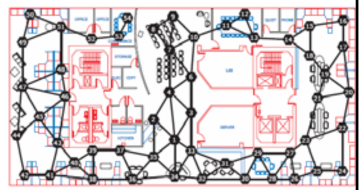
November 1st, 2006

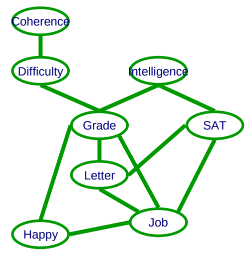
Approximate inference overview

- 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
- There will be a special recitation by Pradeep Ravikumar on more advanced methods a week from Mon.

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

$0 \log 0 = 0$

$\lim_{x \to 0^+} x \log x = 0$

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

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

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

    $p(x) \cdot \log \frac{p(x)}{q(x)} = 0 \cdot \log \frac{0}{q(x)} = 0$ x doesn't appear in distance
  - p(x)≠0 and q(x)=0

    $\underbrace{p(x)}_{=\varepsilon} \log \frac{\overbrace{p(x)}^{=\varepsilon}}{q(x) = 0} = \varepsilon \log \frac{\varepsilon}{0} = +\infty$

    if p thinks p(x) > 0, then q better think q(x) > 0

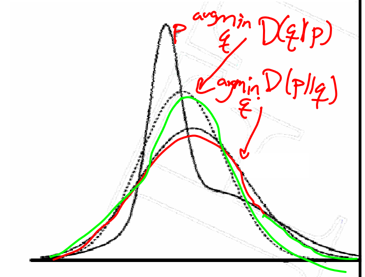
Find simple approximate distribution

- 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 q defines support
  - tends to give overconfident results
  - will be tractable

best

p $\arg\min_q D(q\|p)$

$\arg\min_q D(p\|q)$



p complex - Gaussian w. 2 bumps

q simple - Gaussian w. 1 bump

Back to graphical models

notation drop e

$P(X,e) = \frac{1}{Z}\prod_i \phi_i(C_i)$

- Inference in a graphical model:
  - P(**x**) = $\frac{1}{Z}\prod_i \phi_i(C_i)$
  - want to compute P(X_i|**e**) $\propto P(X_i, e)$
  - our *p*: $P(X) = \frac{1}{Z}\prod_i \phi_i(C_i)$, want $P(X_i)$
- What is the simplest *q*?
  - every variable is independent: q: $Q(X) = \prod_i Q_i(X_i)$
  - mean field approximation
  - can compute any prob. very efficiently

q: X_1 X_2 X_3

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

doesn't depend on q

no role in approx.

- p: $P(X) = \frac{1}{Z}\prod_i \phi_i(C_i)$
- q: $Q(X) = \prod_i Q_i(X_i)$
- D(p||q)= $\sum_x P(x)\log\frac{P(x)}{Q(x)} = \underbrace{\sum_x P(x)\log P(x)}_{-H_p(X)} - \sum_x P(x)\log \underbrace{Q(x)}_{\prod_i Q_i}$

$= -H_p(X) - \sum_x P(x)\log\prod_i Q_i(x_i) = -H_p(X) - \sum_i \sum_x P(x)\log Q_i(x_i)$

$= \sum_{x_i} P(x_i) \underbrace{\sum_{x/x_i} P(x/x_i|x_i)}_{1} \log Q_i(x_i) = \sum_{x_i} P(x_i)\log Q_i(x_i)$ $\sum_i$

need $P(X_i)$ which is what was intractable in the first place

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

$Q(C_i) = \prod_{j: X_j\in C_i} Q_j(X_j)$ Q_i

X_1 X_2 X_3

tractable model

$H_q(X) = \sum_i H_q(X_i)$

- p: $P(x) = \frac{1}{Z}\prod_i \phi_i(C_i)$
- q: $Q(x) = \prod_i Q_i(X_i)$
- D($\cancel{p}q$||$\cancel{q}p$)= $\sum_x q(x)\log\frac{q(x)}{p(x)} = \underbrace{\sum_x q(x)\log q(x)}_{-H_q(X)} - \sum_x q(x)\log p(x)$

$H_q(X) = -\sum_x Q(x)\log\prod_i Q_i(x_i) = -\sum_i \sum_{x_i} Q_i(x_i)\log Q_i(x_i) = \sum_i H_q(X_i)$

$= -\sum_i\sum_x Q(x)\log Q_i(x_i) = -\sum_i \sum_{x_i} \underbrace{Q(x_i)}_{Q_i}\log Q_i(x_i) = \sum_i E_q[\log Q_i(x_i)]$

$\sum_x q(x)\log p(x) = \sum_x q(x)\log\frac{1}{Z}\prod_i \phi_i(C_i) = \sum_i \underbrace{\sum_x Q(x)\log\phi_i(C_i)}_{=\sum_{c_i} Q(c_i)\log\phi_i(c_i)} + \underbrace{\sum_x Q(x)\log Z}_{\log Z\sum_x Q(x)}$

chain rule

$\sum_x Q(x)\log\phi_i(C_i) = \sum_{c_i}\sum_y Q(y|c_i)Q(c_i)\log\phi_i(c_i)$

say $X = C_i \cup Y$

$= \sum_{c_i} Q(c_i)\log\phi_i(c_i)\sum_y \cancel{Q(y|c_i)}$

$= E_Q[\log\phi_i]$ $= +\log Z$

What you need to know so far

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

Announcements

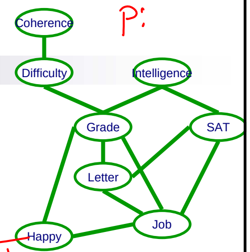
- Tomorrow's recitation
  - Khalid on Variational Methods
- Monday's special recitation
  - Khalid on Dirichlet Processes
    - An exciting way to deal with model selection using graphical models, e.g., selecting (or averaging over) number of clusters in unsupervised learning
    - you even get to see a BN with infinitely many variables

Reverse KL & The Partition Function

Back to the general case

$E_q[f] = \sum_x q(x) f(x)$

P:



- Consider again the defn. of D(q||p):
  - p is Markov net $P_{\mathcal{F}}$ $= \frac{1}{Z}\prod_i \phi_i(c_i)$

$KL(q||P) = \sum_x q(x)\log\frac{q(x)}{P(x)} = -H_q(X) - \sum_x q(x)\log P(x)$

$= -H_q(X) - \sum_i \sum_x q(x)\log\phi_i(c_i) + \log Z$ | $\sum_x q(x)\log\phi_i(c_i) = E_q[\log\phi_i]$

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

$D(q||p) = -H_q(X) - \sum_i E_q[\log\phi_i] + \log Z$

- where energy functional:

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

q:

Understanding Reverse KL, Energy Function & The Partition Function

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

constant

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

- Maximizing Energy Functional ⇔ Minimizing Reverse KL

$\uparrow F[P_F, Q] \Leftrightarrow \downarrow D(Q||P_F)$, $\forall Q\ D(Q||P_F) \geq 0$

- **Theorem**: Energy Function is lower bound on partition function

$\ln Z \geq F[P_F, Q]$ ↑ maximize

$\ln Z = F[P_F, Q]$ iff $Q = P_F$

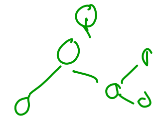
  - Maximizing energy functional corresponds to search for tight lower bound on partition function

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., Graphical model for Q
  - e.g., fully factorized (mean field) $Q(x) = \prod_i Q_i(x_i)$
- Find Q∈Q that maximizes $F[P_{\mathcal{F}}, Q]$
  $\mathcal{F}$ is graphical model for P
- For mean field: $F[P_{\mathcal{F}}, Q] = \sum_{\phi \in \mathcal{F}} E_Q[\log \phi] + \sum_i H_{Q_i}(x_i)$
  $\forall x_i \quad \sum_{x_i} Q_i(x_i) = 1 \quad \boxed{Q_i(x_i) \geq 0 \ \forall x_i}$



Optimization for mean field

$$\max_Q F[P_{\mathcal{F}}, Q] = \sum_{\phi \in \mathcal{F}} E_Q[\ln \phi] + \sum_j H_{Q_j}(X_j), \quad \forall j, \ \sum_{x_j} Q_j(x_j) = 1$$

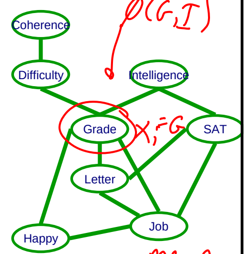
- Constrained optimization, solved via Lagrangian multiplier
  - $\exists\, \lambda$, such that optimization equivalent to:
    $L[Q, \lambda] = \sum_{\phi \in \mathcal{F}} E_Q[\ln \phi] + \sum_j H_{Q_j}(x_j) + \lambda\left(\sum_{x_j} Q_j(x_j) - 1\right)$
  - Take derivative, set to zero
    local minima
    maxima → unstable
    saddle points
- **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

conditional expectation

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



$\phi(G,I)$

$X_i = G$

$E_Q[\ln \phi | x_i] = \sum_{X} Q(X|x_i) \log \phi(C)$

$= \sum_{i \in I} Q(I=i|G=g) \log \phi(G=g, I=i)$

$= \sum_{i \in I} Q(I=i) \log \phi(G=g^{:=t}, I=i)$

$= 0.8 \log 1 + 0.2 \log 2$

$Q(I=i|G=g)$ meanfield

$= Q(I=i)$

$\phi(G,I) =$

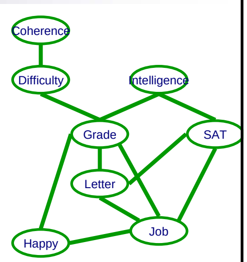
G \ I	t	f
t	1	2
f	3	4

$Q(I) =$

Q	t	f
	0.8	0.2

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

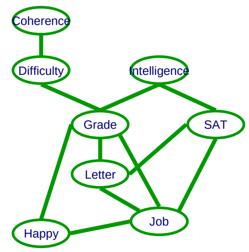
- **Theorem**: The fixed point:

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

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!

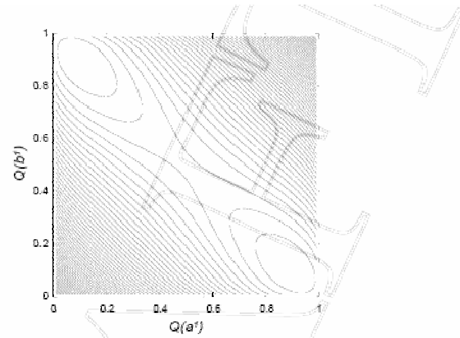
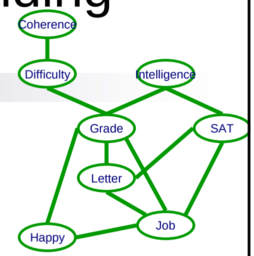


Figure 11.18 An example of a multi-modal mean field energy functional landscape. In this network, $P(a, b) = 0.25 - \epsilon$ if $a \neq b$ and ϵ if $a = b$. The axes correspond to the mean field marginal for A and B and the contours show equi-values of the energy functional.

Very simple approach for finding one stationary point

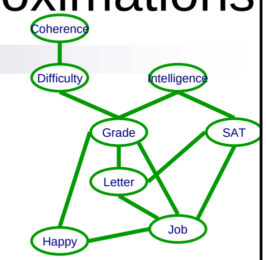


- Initialize Q (e.g., randomly or smartly)
- Set all vars to unprocessed
- Pick unprocessed var X_i
  - update Q_i:

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

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

More general structured approximations

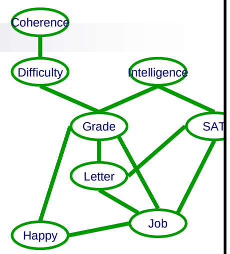


- Mean field very naïve approximation
- Consider more general form for Q
  - assumption: exact inference doable over Q
- **Theorem**: stationary point of energy functional:

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

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 inferece

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

- 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