

Readings:

K&F: 11.3, 11.5

Yedidia et al. paper from the class website

Mean Field and Variational Methods

Loopy Belief Propagation

Graphical Models – 10708

Carlos Guestrin

Carnegie Mellon University

November 8th, 2006

1

Understanding Reverse KL, Energy Function & The Partition Function

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

- Maximizing Energy Functional \Leftrightarrow Minimizing Reverse KL

$$\uparrow F[P_{\mathcal{F}}, Q] \Leftrightarrow \downarrow D(Q || P_{\mathcal{F}}) \quad , \forall Q \quad D(Q || P_{\mathcal{F}}) \geq 0$$

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

$$\ln Z \geq F[P_{\mathcal{F}}, Q] \quad \uparrow \text{maximize}$$

$$\ln Z = F[P_{\mathcal{F}}, Q] \quad \text{iff} \quad Q = P_{\mathcal{F}}$$

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

Structured Variational Approximate Inference

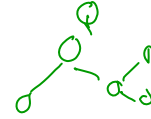
$$\ln Z = F[P_{\mathcal{F}}, Q] + D(Q || P_{\mathcal{F}})$$

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

- 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 $\underline{Q} \in Q$ that maximizes $F[P_{\mathcal{F}}, Q]$

F is graphical model for P

- For mean field: $F(P_{\mathcal{F}}, Q) = \sum_{\phi \in \mathcal{F}} E_Q[\log \phi] + \sum_i H_{Q_i}(x_i)$

$$\forall x_i: \sum_{x_i} Q_i(x_i) = 1 \quad Q_i(x_i) \geq 0 \quad \forall x_i$$

10.708 - ©Carlos Guestrin 2006

3

Optimization for mean field

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

- 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, saddle points, unstable

- **Theorem:** Q is a stationary point of mean field approximation iff for each i :

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

10.708 - ©Carlos Guestrin 2006

4

Understanding fixed point equation

conditional expectation

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

$$E_Q[\ln \phi | x_i] = \sum_x Q(x_i) \log \phi(x_i)$$

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

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

mean field

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

$$Q(I) = \begin{array}{c|cc} & t & f \\ \hline t & 1 & 2 \\ \hline f & 3 & 4 \end{array}$$

10.708 - ©Carlos Guestrin 2006 5

Simplifying fixed point equation

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

10.708 - ©Carlos Guestrin 2006 6

Q_i only needs to consider factors that intersect X_i

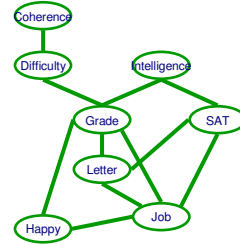
■ **Theorem:** The fixed point:

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

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 $\text{Scope}[\phi_j] = \mathbf{U}_j \cup \{X_i\}$



There are many stationary points!

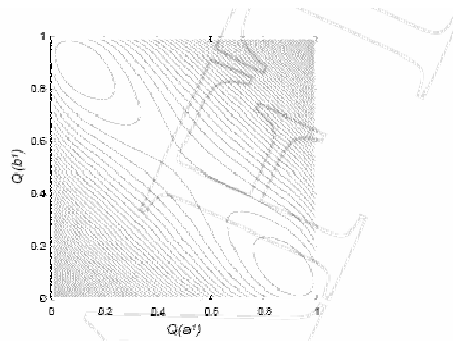
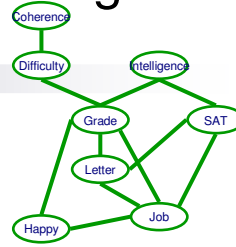


Figure 11.18 An example of a multi-modal mean field energy functional landscape. In this network, $P(a, b) = 0.25 - \epsilon$ if $a \neq b$ and ϵ if $a = b$. The axes correspond to the mean field marginals for A and B and the contours show equi-values of the energy functional.

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



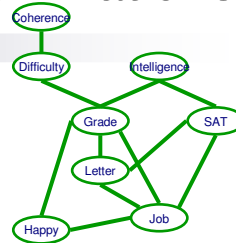
10.708 - ©Carlos Guestrin 2006

9

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:

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



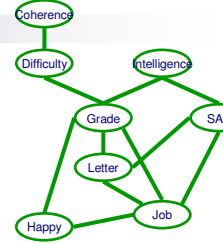
10.708 - ©Carlos Guestrin 2006

10

Computing update rule for general case

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

- Consider one ϕ :



10.708 - ©Carlos Guestrin 2006

11

Structured Variational update requires inference

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

- 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

10.708 - ©Carlos Guestrin 2006

12

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

10.708 – ©Carlos Guestrin 2006

13

Announcements

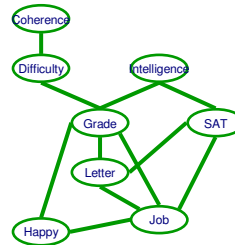
- Tomorrow's recitation
 - Ajit on Loopy BP

10.708 – ©Carlos Guestrin 2006

14

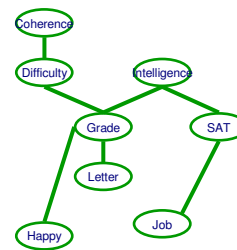
Recall message passing over junction trees

- Exact inference:
 - generate a junction tree
 - message passing over neighbors
 - inference exponential in size of clique



Belief Propagation on Tree Pairwise Markov Nets

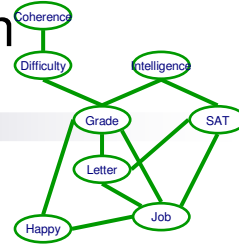
- Tree pairwise Markov net is a tree!!! ☺
 - no need to create a junction tree
- Message passing:



- More general equation:
 - $N(i)$ – neighbors of i in pairwise MN
$$\delta_{i \rightarrow j}(X_j) = \sum_{x_i} \phi_i(x_i) \phi_{ij}(x_i, X_j) \prod_{k \in N(i)-j} \delta_{k \rightarrow i}(x_i)$$

- **Theorem:** Converges to true probabilities:

Loopy Belief Propagation on Pairwise Markov Nets



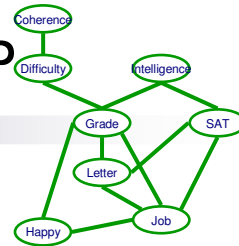
$$\delta_{i \rightarrow j}(X_j) = \sum_{x_i} \phi_i(x_i) \phi_{ij}(x_i, X_j) \prod_{k \in \mathcal{N}(i) - j} \delta_{k \rightarrow i}(x_i)$$

- What if we apply BP in a graph with loops?
 - send messages between pairs of nodes in graph, and hope for the best
- What happens?
 - evidence goes around the loops multiple times
 - may not converge
 - if it converges, usually overconfident about probability values
- But often gives you reasonable, or at least useful answers
 - especially if you just care about the MPE rather than the actual probabilities

10.708 - ©Carlos Guestrin 2006

17

More details on Loopy BP



- Numerical problem:
 - messages < 1 get multiplied together as we go around the loops
 - numbers can go to zero
 - normalize messages to one:

$$\delta_{i \rightarrow j}(X_j) = \frac{1}{Z_{i \rightarrow j}} \sum_{x_i} \phi_i(x_i) \phi_{ij}(x_i, X_j) \prod_{k \in \mathcal{N}(i) - j} \delta_{k \rightarrow i}(x_i)$$

- $Z_{i \rightarrow j}$ doesn't depend on X_j , so doesn't change the answer

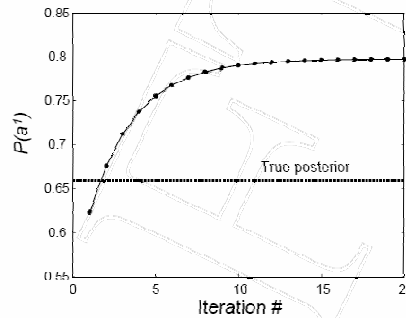
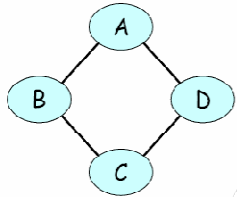
- Computing node "beliefs" (estimates of probs.):

$$\hat{P}(X_i) = \frac{1}{Z_i} \phi_i(X_i) \prod_{k \in \mathcal{N}(i)} \delta_{k \rightarrow i}(X_i)$$

10.708 - ©Carlos Guestrin 2006

18

An example of running loopy BP

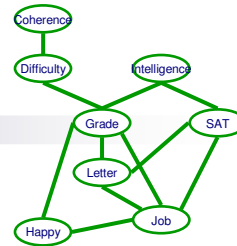


10.708 - ©Carlos Guestrin 2006

19

Convergence

$$\hat{P}(X_i) = \frac{1}{Z_i} \phi_i(X_i) \prod_{k \in \mathcal{N}(i)} \delta_{k \rightarrow i}(X_i)$$



- If you tried to send all messages, and beliefs haven't changed (by much) → converged

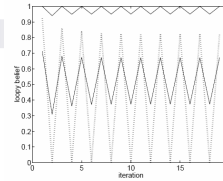
10.708 - ©Carlos Guestrin 2006

20

(Non-)Convergence of Loopy BP

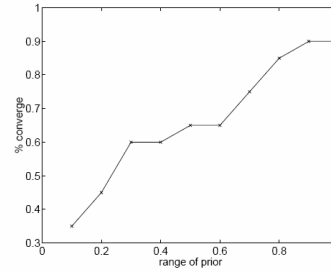
- Loopy BP can oscillate!!!

- oscillations can be small
- oscillations can be really bad!



- Typically,

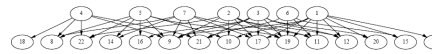
- if factors are closer to uniform, loopy does well (converges)
- if factors are closer to deterministic, loopy doesn't behave well



- One approach to help: damping messages

- new message is average of old message and new one:

- often better convergence
 - but, when damping is required to get convergence, result often bad



graphs from Murphy et al. '99

10.708 - ©Carlos Guestrin 2006

21

Loopy BP in Factor graphs

- What if we don't have pairwise Markov nets?

- Transform to a pairwise MN
- Use Loopy BP on a factor graph



- Message example:

- from node to factor:
- from factor to node:

10.708 - ©Carlos Guestrin 2006

22

Loopy BP in Factor graphs

- From node i to factor j :

- $F(i)$ factors whose scope includes X_i

$$\delta_{i \rightarrow j}(X_i) \propto \prod_{k \in \mathcal{F}(i) - j} \delta_{k \rightarrow i}(X_i)$$

(A) (B) (C) (D) (E)

ABC ABD BDE CDE

- From factor j to node i :

- Scope $[\phi_j] = \mathbf{Y} \cup \{X_i\}$

$$\delta_{j \rightarrow i}(X_i) \propto \sum_{\mathbf{y}} \phi_j(X_i, \mathbf{y}) \prod_{X_k \in \text{Scope}[\phi_j] - X_i} \delta_{k \rightarrow j}(x_k)$$

10.708 - ©Carlos Guestrin 2006

23

What you need to know about loopy BP

- Application of belief propagation in loopy graphs
- Doesn't always converge
 - damping can help
 - good message schedules can help (see book)
- If converges, often to incorrect, but useful results
- Generalizes from pairwise Markov networks by using factor graphs

10.708 - ©Carlos Guestrin 2006

24