ECE 6504: Advanced Topics in Machine Learning
Probabilistic Graphical Models and Large-Scale Learning

Topics
- Markov Random Fields: Inference
  - Approximate: Variational Inference

Readings: KF 11.1,11.2,11.5, Barber 28.1,28.3,28.4

Dhruv Batra
Virginia Tech
Administrativia

• HW3
  – Out 2 days ago
  – Due: Apr 4, 11:55pm
  – Implementation: Loopy Belief Propagation in MRFs

• Project Presentations
  – When: April 22, 24
  – Where: in class
  – 5 min talk
    • Main results
    • Semester completion 2 weeks out from that point so nearly finished results expected
    • Slides due: April 21 11:55pm
Recap of Last Time
Message Passing

• Variables/Factors “talk” to each other via messages:

“I (variable $X_3$) think that you (variable $X_2$):

belong to state 1 with confidence 0.4
belong to state 2 with confidence 10
belong to state 3 with confidence 1.5”
• **Initialization:**
  - Assign each factor $\phi$ to a cluster $\alpha(\phi)$, $\text{Scope}[\phi] \subseteq C_{\alpha(\phi)}$
  - Initialize cluster: $\psi_i^0(C_i) \propto \prod_{\phi: \alpha(\phi) = i} \phi$
  - Initialize messages: $\delta_{j \rightarrow i} = 1$

• While not converged, send messages:
  $\delta_{i \rightarrow j}(S_{ij}) \propto \sum_{C_i - S_{ij}} \psi_i^0(C_i) \prod_{k \in \mathcal{N}(i) - j} \delta_{k \rightarrow i}(S_{ik})$

• **Belief:**
  - On board
Example

• Chain MRF

Compute:

\[ P(X_1 \mid X_5 = x_5) \]

• VE steps on board
Factors Generated

Elimination order:
\[ O = \{C,D,I,H,G,S,L,J\} \]
Cluster graph for VE

• **VE generates cluster tree!**
  (Also called Clique Tree or Junction Tree)
  - One cluster for each factor used/generated
  - Edge $i \to j$, if $f_i$ used to generate $f_j$
  - “Message” from $i$ to $j$ generated when marginalizing a variable from $f_i$
  - Tree because factors only used once

• **Proposition:**
  - “Message” $\delta_{ij}$ from $i$ to $j$
  - $\text{Scope}[\delta_{ij}] \subseteq S_{ij}$
Approximate Inference

• So far: Exact Inference
  – VE & Junction Trees
  – Exponential in tree-width

• There are many many approximate inference algorithms for PGMs
  – You have already seen BP

• Next
  – Variational Inference
  – Connections to BP / Message-Passing
What is Variational Inference?

• A class of methods for approximate inference
  – And parameter learning
  – And approximating integrals basically..

• Key idea
  – Reality is complex
  – Instead of performing approximate computation in something complex
  – Can we perform exact computation in something “simple”?  
  – Just need to make sure the simple thing is “close” to the complex thing.

• Key Problems
  – What is close?
  – How do we measure closeness when we can’t perform operations on the complex thing?

(C) Dhruv Batra
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
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
Example 1

- $p$ = 2D Gaussian with arbitrary co-variance
- $q$ = 2D Gaussian with diagonal co-variance

(argmin_q $KL (p \| q)$)

$p$ = Green; $q$ = Red

(argmin_q $KL (q \| p)$)
Example 2

- $p =$ Mixture of Two Gaussians
- $q =$ Single Gaussian
Back to graphical models

• Inference in a graphical model:
  – $P(x) =$
  – want to compute $P(X_i)$
  – our $p$:

• What is the simplest $q$?
  – every variable is independent:
  – mean field approximation
  – can compute any prob. very efficiently
Variational Approximate Inference

\[ p(x) = \frac{1}{Z} \prod_{(s,t) \in \mathcal{E}} \psi_{st}(x_s, x_t) \prod_{s \in \mathcal{V}} \psi_s(x_s) \]

- Choose a family of approximating distributions which is tractable. The simplest [Mean Field] Approximation:

\[ q(x) = \prod_{s \in \mathcal{V}} q_s(x_s) \]

- Measure the quality of approximations. Two possibilities:

\[ D(p || q) = \sum_x p(x) \log \frac{p(x)}{q(x)} \quad D(q || p) = \sum_x q(x) \log \frac{q(x)}{p(x)} \]

- Find the approximation minimizing this distance
D(p∥q) for mean field – KL the right way

- D(p∥q) =

  - Trivially minimized by setting \( q_i(x_i) = p_i(x_i) \)
  
  - Doesn’t provide a computational method…
Plan for today

- MRF Inference
  - Message-Passing as Variational Inference
    - Mean Field
    - Structured Mean Field
  - (Specialized) MAP Inference
    - Integer Programming Formulation
    - Linear Programming Relaxation
    - Dual Decomposition
D(q||p) for mean field – KL the reverse direction

\[
D(q||p) = \sum_x q(x) \log q(x) - \sum_x q(x) \log p(x)
\]
Reverse KL & The Partition Function

- $D(q || p)$:
  - $p$ is Markov net $P_F$

- **Theorem:**
  $$\log Z = F[p, q] + D(q || p)$$

- Where “Gibbs Free Energy”:
  $$F[p, q] = H_q(\mathcal{X}) + \mathbb{E}_q \left[ \sum_c \log \psi_c(X_c) \right]$$
  $$= H_q(\mathcal{X}) + \mathbb{E}_q \left[ \text{Score}(\mathcal{X}) \right]$$
  $$= H_q(\mathcal{X}) + \sum_c \sum_{x_c} q(x_c) \theta(x_c)$$
Understanding Reverse KL, Free Energy & The Partition Function

\[ \log Z = F[p, q] + D(q||p) \]

\[ F[p, q] = H_q(\mathcal{X}) + \mathbb{E}_q \left[ \sum_c \log \psi_c(X_c) \right] \]

- 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
Mean Field Equations

\[ F[p, q] = H_q(\mathcal{X}) + \mathbb{E}_q \left[ \sum_c \log \psi_c(X_c) \right] \]

\[ H(q) = \sum_{s \in \mathcal{V}} H_s(q_s) = -\sum_{s \in \mathcal{V}} \sum_{x_s} q_s(x_s) \log q_s(x_s) \]

\[ \sum_c \sum_{x_c} q_c(x_c) \theta(x_c) = \sum_i \sum_{x_i} q_i(x_i) \theta_i(x_i) + \sum_{(i,j) \in E} \sum_{x_i} \sum_{x_j} q_i(x_i) q_j(x_j) \theta_{ij}(x_i, x_j) \]

- Add Lagrange multipliers to enforce \( \sum_{x_s} q_s(x_s) = 1 \)

- Taking derivatives and simplifying, we find a set of fixed point equations:
  \[ q_i(x_i) \propto \psi_i(x_i) \prod_{j \in N(i)} \exp \left\{ \sum_{x_j} \theta_{ij}(x_i, x_j) q_j(x_j) \right\} \]

- Updating one marginal at a time gives convergent coordinate descent
Mean Field versus Belief Propagation

\[ p(x) = \frac{1}{Z} \prod_{(s,t) \in \mathcal{E}} \psi_{st}(x_s, x_t) \prod_{s \in \mathcal{V}} \psi_s(x_s) \]

\[ q_t(x_t) \propto \psi_t(x_t) \prod_{u \in \Gamma(t)} m_{ut}(x_t) \]

**BP:**

**MF:**

**Big implications from small changes:**

- **Belief Propagation:** Produces exact marginals for any tree, but for general graphs no guarantees of convergence or accuracy
- **Mean Field:** Guaranteed to converge for general graphs, always lower-bounds partition function, but approximate even on trees
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 $\epsilon$ 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.
CRF models in multi-class image segmentation

\[ E(x) = \sum_i \psi_u(x_i) + \sum_i \sum_{j \in N_i} \psi_p(x_i, x_j) \]

- MAP inference in conditional random field
- Unary term
  - From classifier
  - TextonBoost [Shotton et al. 09]
- Pairwise term
  - Consistent labeling
Adjacency CRF models

\[ E(x) = \sum_i \psi_u(x_i) + \sum_i \sum_{j \in N_i} \psi_p(x_i, x_j) \]

- Efficient inference
  - 1 second for 50,000 variables
- Limited expressive power
- Only local interactions
- Excessive smoothing of object boundaries
  - Shrinking bias
Adjacency CRF models

\[ E(x) = \sum_i \psi_u(x_i) + \sum_i \sum_{j \in N_i} \psi_p(x_i, x_j) \]

- Efficient inference
  - 1 second for 50,000 variables
- Limited expressive power
- Only local interactions
- Excessive smoothing of object boundaries
  - Shrinking bias
Adjacency CRF models

\[ E(x) = \sum_i \psi_u(x_i) + \sum_i \sum_{j \in N_i} \psi_p(x_i, x_j) \]

- Efficient inference
  - 1 second for 50’000 variables
- Limited expressive power
- Only local interactions
- Excessive smoothing of object boundaries
  - Shrinking bias
Fully connected CRF

\[ E(x) = \sum_i \psi_u(x_i) + \sum_i \sum_{j>i} \psi_p(x_i, x_j) \]

- Every node is connected to every other node
  - Connections weighted differently
Fully connected CRF

\[ E(x) = \sum_{i} \psi_u(x_i) + \sum_{i} \sum_{j > i} \psi_p(x_i, x_j) \]

- Long-range interactions
- No more shrinking bias
Fully connected CRF

\[ E(x) = \sum_i \psi_u(x_i) + \sum_i \sum_{j>i} \psi_p(x_i, x_j) \]

- Long-range interactions
- No more shrinking bias
Fully connected CRF

\[ E(x) = \sum_{i} \psi_u(x_i) + \sum_{i} \sum_{j>i} \psi_p(x_i, x_j) \]

- Region-based [Rabinovich et al. 07, Galleguillos et al. 08, Toyoda & Hasegawa 08, Payet & Todorovic 10]
  - Tractable up to hundreds of variables
- Pixel-based
  - Tens of thousands of variables
    - Billions of edges
  - Computationally expensive
Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials

- Inference in 0.2 seconds
  - 50'000 variables
  - MCMC inference: 36 hrs
- Pairwise potentials: linear combinations of Gaussians
Inference

Find the most likely assignment (MAP)

\[ \hat{x} = \arg\max_x P(x) \quad \text{where} \quad P(x) = \exp(-E(x)) \]

Mean field approximation

- Find \( Q(x) = \prod_i Q(x_i) \) close to \( P(x) \) in terms of KL-divergence \( D(Q\|P) \)
- \( \hat{x}_i \approx \arg\max_{x_i} Q(x_i) \)
Results: MSRC

**MSRC dataset**
- 591 images
- 21 classes

<table>
<thead>
<tr>
<th></th>
<th>Time</th>
<th>Global</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unary</td>
<td>-</td>
<td>84.0</td>
<td>76.6</td>
</tr>
<tr>
<td>Grid CRF</td>
<td>1s</td>
<td>84.6</td>
<td>77.2</td>
</tr>
<tr>
<td>FC CRF</td>
<td>0.2s</td>
<td>86.0</td>
<td>78.3</td>
</tr>
</tbody>
</table>
Summary

- **Fully connected CRF model**
  - Pairwise terms: linear combination of Gaussians

- **Efficient inference**
  - Linear in number of variables
  - Independent of number of pairwise terms
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