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

Topics
- Markov Random Fields: MAP Inference
  - Max-Product Message Passing
  - Integer Programming, LP formulation
  - Dual Decomposition

Readings: KF 13.1-5, Barber 5.1,28.9

Dhruv Batra
Virginia Tech
Administrativia

• HW1 Solutions
  – Released
  – Grades almost done too

• 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”
Generalized BP

• Initialization:
  – Assign each factor $\phi$ to a cluster $\alpha(\phi)$, Scope[$\phi$] $\subseteq C_{\alpha(\phi)}$
  – Initialize cluster: $\psi^0_i(C_i) \propto \prod_{\phi: \alpha(\phi) = i} \phi$
  – Initialize messages: $\delta_{j \rightarrow i} = 1$

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

• Belief:
  – On board
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?
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…
\(D(q \| p)\) 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”:
  \[
  F[p, q] = H_q(\mathcal{X}) + \mathbb{E}_q \left[ \sum_c \log \psi_c(X_c) \right]
  \]
  \[
  = H_q(\mathcal{X}) + \mathbb{E}_q [\text{Score}(\mathcal{X})]
  \]
  \[
  = H_q(\mathcal{X}) + \sum_c \sum_{x_c} q(x_c) \theta(x_c)
  \]
Understanding Reverse KL, Free Energy & The Partition Function

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

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

- Maximizing Energy Functional \(\Leftrightarrow\) 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

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

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

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

• 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
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
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
Plan for today

• MRF Inference
  – (Specialized) MAP Inference
    • Integer Programming Formulation
    • Linear Programming Relaxation
    • Dual Decomposition
Possible Queries

• Evidence: $E=e$ (e.g. $N=t$)
• Query variables of interest $Y$

• Conditional Probability: $P(Y \mid E=e)$
  – E.g. $P(F,A \mid N=t)$
  – Special case: Marginals $P(F)$

• Maximum a Posteriori: $\text{argmax } P(\text{All variables} \mid E=e)$
  – $\text{argmax}_{\{f,a,s,h\}} P(f,a,s,h \mid N = t)$

• Marginal-MAP: $\text{argmax}_y P(Y \mid E=e)$
  – $= \text{argmax}_y \sum_o P(Y=y, O=o \mid E=e)$
MAP Inference

$S(y) = \sum_{i \in V} \theta_i(y_i) + \sum_{(i,j) \in E} \theta_{ij}(y_i, y_j)$

$P(y) = \frac{1}{Z} e^{S(y)}$

Node Scores / Local Rewards

$y_1 \quad y_2 \quad \ldots \quad y_n$

Most Likely Assignment

$\arg\max_y S(y)$

$P(y)$

Edge Scores / Distributed Prior

$\text{Person}$

$\text{Table}$

$\text{Plate}$

(C) Dhruv Batra
Example

- Chain MRF

- Max-Product VE steps on board
Loopy BP on Pairwise Markov Nets

\[
\delta_{i\rightarrow j}(y_j) = \sum_{y_i} \phi_i(y_i) \phi_{ij}(y_i, y_j) \prod_{k \in N(i) - j} \delta_{k\rightarrow i}(y_i)
\]
MAP in Pairwise MRFs

- Over-Complete Representation

\[
\theta = \begin{bmatrix}
\theta_1(1) & \ldots & \theta_1(k) \\
\theta_n(1) & \ldots & \theta_n(k) \\
\theta_{12}(1,1) & \ldots & \theta_{12}(k,k) \\
\theta_{n-1,n}(1,1) & \ldots & \theta_{n-1,n}(k,k)
\end{bmatrix}
\]

\[
\mu = \begin{bmatrix}
\mu_1(1) & \ldots & \mu_1(k) \\
\mu_n(1) & \ldots & \mu_n(k)
\end{bmatrix}
\]

- Over-Complete Representation

\[
\mathbf{x} = \begin{bmatrix}
x_1 \\
\vdots \\
x_n \\
(x_1, x_2) \\
\vdots \\
(x_{n-1}, x_n)
\end{bmatrix}
\]

\[
x_i = \begin{bmatrix}
\mu_i(s) \\
\vdots \\
k \times 1
\end{bmatrix}
\]

\[
\begin{bmatrix}
1 \\
0 \\
0 \\
0 \\
0
\end{bmatrix}
\]

\[
\begin{bmatrix}
1 & 0 \\
1 & 0 \\
0 & 0 \\
0 & 0
\end{bmatrix}
\]

\[
x_i = 1 \\
x_i = 2
\]
MAP in Pairwise MRFs

\[ G = (\mathcal{V}, \mathcal{E}) \]

- Over-Complete Representation

\[ \theta = \begin{bmatrix} \theta_1(1) \ldots \theta_1(k) \\ \vdots \\ \theta_n(1) \ldots \theta_n(k) \\ \theta_{12}(1, 1) \ldots \theta_{12}(k, k) \\ \vdots \\ \theta_{n-1,n}(1, 1) \ldots \theta_{n-1,n}(k, k) \end{bmatrix} \]

\[ \mu_x = \begin{bmatrix} \mu_1(1) \ldots \mu_1(k) \\ \vdots \\ \mu_n(1) \ldots \mu_n(k) \\ \mu_{12}(1, 1) \ldots \mu_{12}(k, k) \\ \vdots \\ \mu_{n-1,n}(1, 1) \ldots \mu_{n-1,n}(k, k) \end{bmatrix} \]

\[ S(x) = \theta \cdot \mu_x \]

(C) Dhruv Batra
MAP in Pairwise MRFs

• Integer Program

\[
\max_{\mu} \quad \theta^T \mu
\]

\[
\mu_i(s) \in \{0, 1\}
\]
\[
\mu_{ij}(s, t) \in \{0, 1\}
\]

\[
\sum_s \mu_i(s) = 1
\]

\[
\sum_{s,t} \mu_{i,j}(s, t) = 1
\]

\[
\sum_s \mu_{ij}(s, t) = \mu_j(t)
\]
MAP in Pairwise MRFs

- MAP Integer Program

\[
\begin{align*}
\max_{\mu} & \quad \theta^T \mu \\
\text{s.t.} & \quad A\mu = b \\
& \quad \mu(\cdot) \in \{0, 1\}
\end{align*}
\]
MAP in Pairwise MRFs

• MAP Linear Program

\[
\max_{\mu} \quad \theta^T \mu \\
\text{s.t.} \quad A\mu = b \\
\mu(\cdot) \in [0, 1]
\]

• Properties
  – If LP-opt is integral, MAP is found
  – LP always integral for trees
  – Efficient message-passing schemes for solving LP
MAP in Pairwise MRFs

• Compare MAP LP to Variational Inference
  – On board
  – Difference in entropy term (objective)
  – Family of Q distributions
MAP in Pairwise MRFs

- MAP Linear Program

\[
\begin{align*}
\max_{\mu} & \quad \theta^T \mu \\
\text{s.t.} & \quad A \mu = b \\
& \quad \mu(\cdot) \in [0, 1]
\end{align*}
\]

\[
A = \begin{pmatrix}
\vdots \\
\end{pmatrix} \quad O(|\mathcal{V}|)
\]

Off-the-shelf solvers
- CPLEX
- Mosek
- etc

(C) Dhruv Batra
LP Relaxation

• Block Co-ordinate / Sub-gradient Descent on Dual

\[ A = O(|\mathcal{E}|) \]

\[ O(|\mathcal{E}|) \]
LP Relaxation

• Block Co-ordinate / Sub-gradient Descent on Dual

\[ A = \begin{pmatrix} \lambda_{ij \rightarrow j}^{(t)} \\ \lambda_{ji \rightarrow i}^{(t)} \end{pmatrix} \]

\[ O(|\mathcal{E}|) \]

\[ O(|\mathcal{E}|) \]
LP Relaxation

- Block Co-ordinate / Sub-gradient Descent on Dual

\[
A = \begin{pmatrix}
& \lambda_{ij \rightarrow j}^{(t+1)} \\
& \lambda_{ji \rightarrow i}^{(t+1)}
\end{pmatrix} \quad O(|\mathcal{E}|)
\]

Distributed Message-Passing

Still inefficient!