Variational Inference

Amr Ahmed

Nov. 6\textsuperscript{th} 2008
Outline

• Approximate Inference
• Variational inference formulation
  – Mean Field
    • Examples
  – Structured VI
    • Examples
Approximate Inference

- Exact inference is \textit{exponential} in clique size
- Posterior is \textit{highly peaked}
  - Can be approximated by a \textit{simpler} distribution
- Formulate inference as an optimization problem
  - Define an objective: how good is $Q$
  - Define a \textit{family of simpler distributions} to search over
  - Find $Q^*$ that best approximate $P$
Approximate Inference

- Exact inference is exponential in clique size
- Posterior inference is highly peaked
  - Can be approximated by a simpler distribution
- Formulate inference as an optimization problem
  - Define an objective: how good is $Q$
  - Define a family of simpler distributions to search over
  - Find $Q^*$ that best approximate $P$
- Today we will cover variational Inference
  - Just a possible way of such a formulation
- There are many other ways
  - Variants of loopy BP (later in the semester)
What is Variational Inference?

- Tractable Posterior Family
  - Set CPT (factors)
  - Get distribution, Q

\[ Q^* = \text{Min}_{Q \in Q} KL(Q \| P) \]

- Fully instantiated distribution \( Q^* \)
- Enables exact Inference
  - Has low tree-width
  - Clique tree, variable elimination, etc.

- Posterior “hard” to compute
- We know the CPTs (factors)
VI Questions

• Which family to choose
  – Basically we want to remove some edges
    • Mean field: fully factorized
    • Structured: Keep tractable sub-graphs

• How to carry the optimization

\[ Q^* = \min_{Q \in Q} KL(Q\|P) \]

• Assume P is a Markov network
  – Product of factors (that is all)
Outline

• Approximate Inference
• Variational inference formulation
  – Mean Field
    • Examples
  – Structured VI
    • Examples
Mean-field Variational Inference

• Tractable Posterior **Family**
  • Set CPT (factors)
  • Get distribution, Q

\[
Q^* = \text{Min}_{Q \in Q} KL(Q \| P)
\]

• Fully instantiated distribution \( Q^* \)
• Enables exact Inference

• Posterior “hard” to compute
• We know the CPTs (factors)
D(q || p) for mean field – 
KL the reverse direction: cross-entropy term

- p: $\frac{1}{Z} \prod_j \phi_i(c_i)$
- q: $\prod_j Q_j(x_j)$

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

\[
= \sum_x q(x) \log p(x) - \sum_x q(x) \log \frac{1}{Z} \prod_i \phi_i(c_i)
\]

\[
= \sum_x q(x) \log \phi_i(c_i) - \sum_x q(x) \log Z
\]

\[
= \sum_{c_i} q(c_i) \log \phi_i(c_i)
\]

For mean fields:

\[
q(c_i) = \prod_{x_j \in c_i} Q_j(x_j)
\]

(As long as $c_i$ not too large)
The Energy Functional

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

- Where energy functional:
  \[
  F[P_\mathcal{F}, Q] = \sum_{\phi \in \mathcal{F}} E_Q[\ln \phi] + H_Q(\mathcal{X})
  \]

- Our problem now is
  \[
  Q^* = \max_{Q \in \mathcal{Q}} F[P_F, Q]
  \]
  \[
  \sum_{x_i} Q(x_i) = 1
  \]

You get

\( \ln Z \leq F[P_F, Q] \quad \forall Q \)

Maximizing \( F[P,Q] \) **tighten** the bound
And gives better prob. estimates
Our problem now is

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

Tractable by construction

\[ Q^* = \max_{Q \in \mathcal{Q}} \frac{F[P_F, Q]}{\sum_{x_1} Q(x_1) = 1} \]

Theorem: \( Q \) is a stationary point of mean field approximation iff for each \( j \):

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

• Approximate Inference
• Variational inference formulation
  – Mean Field
    • Examples
  – Structured VI
    • Examples
Given the factors in $P$, we want to get the factors for $Q$.

- Iterative procedure. Fix all $q_{-i}$, compute $q_i$ via the above equation
- Iterate until you reach a fixed point
$Q_i(x_i) = \frac{1}{Z_i} \exp \left\{ \sum_{\phi \in \mathcal{F}} E_Q[\ln \phi \mid x_i] \right\}$

$P(X) \propto \prod_{i, j \in E} \phi(X_i, X_j)$

Example

$Q_i(x_i) \propto \exp \left\{ \sum_{\phi_i : X_i \in \text{scope} [\phi_i]} E_Q(U_{\phi_i \setminus \{x_i\}})[\ln \phi_i(x_i, U_{\phi_i})] \right\}$

$q_i(x_i) \propto \exp \left\{ E_Q(\{x_1, x_2\} \setminus \{x_i\})[\ln \phi(x_1, x_2)] + E_Q(\{x_1, x_5\} \setminus \{x_i\})[\ln \phi(x_1, x_5)] \right\}$
\[ Q_i(x_i) = \frac{1}{Z_i} \exp \left\{ \sum_{\phi \in \mathcal{F}} E_Q[\ln \phi | x_i] \right\} \]

Example

**Q**: fully factorized MN

**P**: Pairwise MN

\[ P(X) \propto \prod_{i, j \in E} \phi(X_i, X_j) \]

\[ q_i(x_i) \propto \exp \left\{ \sum_{\phi_i : x_i \in \text{scope} [\phi_i]} E_{Q(U_\phi - \{x_i\})}[\ln \phi_i(x_i, U_\phi)] \right\} \]

\[ q_i(x_1) \propto \exp \left\{ E_{q(x_2)}[\ln \phi(x_1, x_2)] + E_{q(x_5)}[\ln \phi(x_1, x_5)] \right\} \]

\[ \propto \exp \left\{ \sum_{x_2} q_2(x_2) \ln \phi(x_1, x_2) + \sum_{x_5} q_5(x_5) \ln \phi(x_1, x_5) \right\} \]
\[ Q_i(x_i) = \frac{1}{Z_i} \exp \left\{ \sum_{\phi \in \mathcal{F}} E_Q[\ln \phi | x_i] \right\} \]

**Example**

**Q**: fully factorized MN

\[ P(X) \propto \prod_{i, j \in E} \phi(X_i, X_j) \]

\[ q_i(x_i) \propto \exp \left\{ \sum_{\phi_i: X_i \in \text{scope } [\phi_i]} E_{Q(U_{\phi \setminus \{X_i\})}}[\ln \phi_i(x_i, U_{\phi})] \right\} \]

In your homework
Example

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

-\( P(X) = \prod_{i, j \in E} \phi(X_i, X_j) \)

Intuitively

- \( q(X_6) \) get to be tied with \( q(x_i) \) for all \( x_i \) that appear in a factor with it in \( P \)
  - i.e. fixed point equations are tied according to \( P \)
- What \( q(x_6) \) gets is expectations under these \( q(x_i) \) of how the factor looks like
- can be somehow interpreted as message passing as well (but we won’t cover this)

\[ q_i(x_i) \propto \exp \left\{ \sum_{\phi_i : X_i \in \text{scope } [\phi_i]} E_Q(U_{\phi_i \setminus \{X_i\}})[\ln \phi_i(x_i, U_{\phi_i})] \right\} \]
Example

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

P: Pairwise MN

\[ P(X) = \prod_{i, j \in E} \phi(X_i, X_j) \]

Q: fully factorized MN

Intuitively

- \( q(X_6) \) get to be tied with \( q(x_i) \) for all \( x_i \) that appear in a factor with it in \( P \)
  - i.e. fixed point equations are tied according to \( P \)
- What \( q(x_6) \) gets is expectations under these \( q(x_i) \) of how the factor looks like
- can be somehow interpreted as message passing as well (but we won’t cover this)

\[ q_i(x_i) \propto \exp \left\{ \sum_{\phi_i : \text{scope } [\phi_i]} E_Q(U_{\phi_i} - \{ x_i \}) \left[ \ln \phi_i(x_i, U_{\phi_i}) \right]_i \right\} \]
Outline

• Approximate Inference
• Variational inference formulation
  – Mean Field
    • Examples
  – Structured VI
    • Examples
• Inference in LDA (time permits)
Structured Variational Inference

• Tractable Posterior Family
  • Set CPT (factors)
  • Get distribution, Q

\[ Q^* = \min_{Q \in \mathcal{Q}} KL(Q\|P) \]

• Fully instantiated distribution \( Q^* \)
• Enables exact Inference
  • Has low tree-width
  • You don’t have to remove all edges
  • Just keep tractable sub graphs: trees, chain

• Posterior “hard” to compute
• We know the CPTs (factors)
Structured Variational Inference

\[ Q(X) \propto \prod_{i, j \in E_q} \psi_{i,j}(X_i, X_j) \]

\[ P(X) \propto \prod_{i, j \in E} \phi(X_i, X_j) \]

**Goal**

Given factors in P, get factors in Q that minimize energy functional

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

Still tractable

By construction
Fixed point Equations

Q: tractably factorized MN

Q(X) \propto \prod_{i, j \in E_q} \psi_{i,j}(X_i, X_j)

P: Pairwise MN

P(X) \propto \prod_{i, j \in E} \phi(X_i, X_j)

\psi_j(c_j) \propto \exp \left[ \sum_{\phi \in A_j} E_Q[\ln \phi | c_j] - \sum_{\psi_k \in B_j} E_Q[\ln \psi_k | c_j] \right]

Factors that scope() is not independent of C_j in Q
Fixed point Equations

\[ \text{Q: tractably factorized MN} \]

\[ Q(X) \propto \prod_{i, j \in E_q} \psi_{i, j}(X_i, X_j) \]

\[ P(X) \propto \prod_{i, j \in E} \phi(X_i, X_j) \]

\[ \psi_j(c_j) \propto \exp \left[ \sum_{\phi \in A_j} E_Q[\ln \phi | c_j] - \sum_{\psi_k \in B_j} E_Q[\ln \psi_k | c_j] \right] \]

Factors that `scope()` is not independent of \( C_j \) in \( Q \)
**Fixed point Equations**

\[ \psi_j(c_j) \propto \exp \left[ \sum_{\phi \in A_j} E_Q[\ln \phi | c_j] - \sum_{\psi_k \in B_j} E_Q[\ln \psi_k | c_j] \right] \]

**Q : tractably factorized MN**

**P: Pairwise MN**

What are the factors in P with scope not separated from \( x_1, x_5 \) in **Q**

- First, get variables that cannot be separated from \( x_1 \) and \( x_5 \) in **Q**
- Second collect the factors they appear in in **P**
Fixed point Equations

Q: tractably factorized MN

\[
\psi_j(c_j) \propto \exp \left[ \sum_{\phi \in A_j} E_Q[\ln \phi | c_j] - \sum_{\psi_k \in B_j} E_Q[\ln \psi_k | c_j] \right]
\]

P: Pairwise MN

What are the factors in P with scope \textbf{not separated} from x_1, x_5 in Q

- First, get variables that cannot be separated from x_1 and x_5 in Q
- Second collect the factors they appear in in P
- Expectation is taken over Q(slope[factor] \mid c_j)
P: Pairwise MN

Q: tractably factorized MN

Fixed point Equations

\[ \psi_j(c_j) \propto \exp \left[ \sum_{\phi \in A_j} E_Q[\ln \phi | c_j] - \sum_{\psi_k \in B_j} E_Q[\ln \psi_k | c_j] \right] \]

\[ Q(X) \propto \prod_{i, j \in E_q} \psi_{i,j}(X_i, X_j) \]

\[ B = \psi_{59} \]

\[ A = \phi_{15} \phi_{59} \phi_{12} \phi_{56} \phi_{9,10} \]

\[ \psi_j(c_j) \propto \exp \left[ \sum_{\phi \in A_j} E_Q[\ln \phi | c_j] - \sum_{\psi_k \in B_j} E_Q[\ln \psi_k | c_j] \right] \]

\[ E_Q[\ln \phi(x_5, x_9) | x_1, x_5] = E_{Q(x_5|x_1,x_5)}[\ln \phi(x_5, x_9)] \]
$\psi_j(c_j) \propto \exp \left[ \sum_{\phi \in A_j} E_Q[\ln \phi | c_j] - \sum_{\psi_k \in B_j} E_Q[\ln \psi_k | c_j] \right]$ 

**Q**: tractably factorized MN

**P**: Pairwise MN

$Q(X) \propto \prod_{i, j \in E_q} \psi_{i,j}(X_i, X_j)$

$P(X) \propto \prod_{i, j \in E} \phi(X_i, X_j)$

$A = \psi_{59}$

$B = \phi_{15} \phi_{59} \phi_{12} \phi_{56} \phi_{9,10}$

$E_Q[\ln \psi_{5,9}(x_5, x_9) | x_1, x_5] = E_{Q(x_5|x_1,x_3)}[\ln \psi(x_5, x_9)]$
\[ Q_j(c_j) \propto \exp \left[ \sum_{\phi \in A_j} E_Q[\ln \phi | c_j] - \sum_{\psi_k \in B_j} E_Q[\ln \psi_k | c_j] \right] \]

**Q**: tractably factorized MN

\[
Q(X) \propto \prod_{i, j \in E_q} \psi_{i, j}(X_i, X_j)
\]

\[
E_Q[\ln \phi_{9,10}(x_9, x_{10}) | x_1, x_5] = E_{Q(x_9, x_{10} | x_1, x_5)}[\ln \phi_{9,10}(x_9, x_{10})]
\]

\[
Q(x_9, x_{10} | x_1, x_5) = Q(x_9 | x_1, x_5)Q(x_{10})
\]

**P**: Pairwise MN

\[
P(X) \propto \prod_{i, j \in E} \phi(X_i, X_j)
\]

**Fixed point Equations**
Some intuition

$$Q(X) \propto \prod_{i, j \in E_q} \psi_{i,j}(X_i, X_j)$$

$$P(X) \propto \prod_{i, j \in E} \phi(X_i, X_j)$$

-What have we picked from P when dealing with the first factor (X1,x5)?
  -Factors that interact with (x1,x5) in both P and Q
  - Fixed point equations are tied in the same way these variables are tied in P and Q
- can be also interpreted as passing message over Q of expectations on these factors
- Q: how to compute these madrigals in Q efficiently? Homework!

Q: **tractably** factorized MN

\[ \psi_j(c_j) \propto \exp \left[ \sum_{\phi \in A_j} E_Q[\ln \phi|c_j] - \sum_{\psi_k \in B_j} E_Q[\ln \psi_k|c_j] \right] \]
Variational Inference in Practice

• Highly used
  – Fast and efficient
  – Approximation quality might be not that good
    • Always peaked answers (we are using the wrong KL)
    • Always captures modes of P
    • Spread of P is usually lost
  – Famous applications of VI
    • VI for topic models (LDA)
      – You can derive LDA equations in few lines using the fully factorized update equations (exercise)
      – We might go over it in a special recitations but try it if you are doing a project in topic models
    • Involved temporal models
    • Many more...