Mean Field / Variational Approximations

Presented by Jose Nuñez

10/24/05

Outline

• Introduction
• Mean Field Approximation
• Structured Mean Field
• Weighted Mean Field
• Variational Methods
Introduction

Problem:

• We have distribution $P(x)$ but inference is hard to compute.

Previous solutions:

• Approximate energy functional: Bethe, Kikuchi

New idea:

• Directly optimize the energy functional introducing a distribution $Q(x)$ defined on the same domain of variables as $P$ which incorporates some constraints.

• Objective: We want to find $Q(x)$ which is the best approximation of $P(x)$ and use $Q(x)$ to make inferences.

• Find $Q \in Q$ that minimizes $F(P,Q)$
Mean Field Approximation

Assumptions:

- $Q(x)$ is our mean field approximation.
- Variables in the $Q$ distribution are independent variables $X_i$.
- In the standard mean field approach, $Q$ is completely factorized:

$$Q(x) = \prod Q_i(x_i)$$

What happens when we apply mean field?

![Diagram showing the transformation from $P(x)$ to $Q(x)$]

$$P(x) \quad Q(x)$$
Mean Field Approximation

\[
F(P, Q) = -\sum_{\phi \in F} \sum_{x_\phi} Q(x_\phi) \log \phi(x_\phi) + \sum_x Q(x) \log Q(x)
\]
\[
Q(x) = \prod_i Q_i(x_i)
\]
\[
E(P, Q) = -\sum_{\phi \in F} \sum_{x_\phi} Q(x_\phi) \log \phi(x_\phi) = -\sum_{\phi \in F} \sum_{x_\phi} \left( \prod_{i \in \phi} Q_i(x_i) \right) \log \phi(x_\phi)
\]
\[
H(Q) = -\sum_x Q(x) \log Q(x) = -\sum_x \left( \prod_{i \in x} Q_i(x_i) \right) \log \prod_{i \in x} Q_i(x_i)
\]
\[
= -\sum_x \left( \prod_{i \in x} Q_i(x_i) \right) \log \left( \sum_{x_i} \prod_{i \in x} Q_i(x_i) \right)
\]
\[
= -\sum_i \sum_{x_i} Q(x_i) \log Q(x_i)
\]
\[
= \sum_i H_{Q_i}(x_i)
\]

Task: find \( \sum_{x_i} Q(x_i) = 1 \) minimizing \( F(P, Q) \)

Solving: build a Lagrangian, differentiate and set to 0!
Mean Field Approximation

The distribution $Q(x_i)$ is locally optimal solution given $Q(x_1), \ldots, Q(x_{i-1}), Q(x_{i+1}), \ldots, Q(x_n)$, if:

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

Where $Z_i$ is a local normalizing constant and $E_Q[\ln \phi | x_i]$ is the conditional expectation given the value $x_i$.

Mean Field Approximation

Locality:

- Only local operations are needed for iteration of the MF-equations.
- In other words, only neighboring variables are needed.

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

where $U_\phi = \text{Scope} [\phi]$

Calculation of $Q(X_i)$ depends only on clusters $X_i$ belongs to.
**Mean Field Approximation**

**Solution:** Iterate mean field equations

\[
Q(x_i) = \frac{1}{Z_i} \exp \left\{ \sum_{\phi, x \in \text{Scope}[\phi]} E_{\phi} \left[ \ln \phi(U_{\phi}, x_i) \right] \right\}
\]

- Converge to a fixed point.

**Problem:** convergence to a local optima.

---

**Mean Field Approximation**

Haft et al. paper:

- Optimize the KL divergence instead of the free energy

\[
D(Q \| P) = E_{\phi} \left( \log \frac{Q(x)}{P(x)} \right)
\]

\[
D(Q \| P) = E_{\phi} \left( \log Q(x) \right) - E_{\phi} \left( \log P(x) \right)
\]

\[
D(Q \| P) = E_{\phi} \left( \log Q(\bar{X}_i) \right) - E_{\phi} \left( \log P(\bar{X}_i) \right)
\]

\[
+ E_{\phi} \left( \log Q(X_i) \right) - E_{\phi} \left( \log P(X_i | \bar{X}_i) \right)
\]

Assume: \( P(X) = P(X_i | \bar{X}_i) P(\bar{X}_i) \)

\[
D(Q \| P) = E_{\phi(x_i)} \left( \log Q(\bar{X}_i) \right) - E_{\phi(x_i)} \left( \log P(\bar{X}_i) \right)
\]

\[
+ E_{\phi(x_i)} \left( \log Q(X_i) \right) - E_{\phi(x_i)} \left( \log P(X_i | \bar{X}_i) \right)
\]
Mean Field Approximation

Haft et al. paper:
- Optimize the KL divergence instead of the free energy

\[
D(Q \mid P) = E_{Q(X_i)}(\log Q(X_i)) - E_{Q(X_i)}(\log P(X_i)) + E_{Q(X_i)}(\log Q(X_i)) - E_{Q(X_i)}(\log P(X_i \mid X_i))
\]

\[
\frac{\partial D(Q \mid P)}{\partial Q(X_i)} = \frac{\partial}{\partial Q(X_i)} E_{Q(X_i)}(\log Q(X_i)) - E_{Q(X_i)}(\log P(X_i \mid X_i))
\]

Subject to \( \sum_{x_i} Q(X_i) = 1 \)

Mean Field Approximation

Haft et al. paper:

\[
Q(X_i) \propto \exp \left( E_{Q(X_i)}(\log P(X_i \mid X_i)) \right) = \exp \left( \log P(X_i \mid X_i) \right)_{Q(X_i)} \quad \text{MF-equation}
\]

Locality:

\[
Q(x_i) \propto \exp \left( \log P(x_i \mid M_i) \right)_{Q(M_i)} \quad \text{MF-equation simplified}
\]

where \( M \) is the Markov boundary.
Mean Field Approximation

Algorithm:

Procedure Mean-Field ( \\
\( F \) \quad \text{// factors that define } P_F \\
Q_0 \quad \text{// initial choice of } Q \\
) \\
\begin{align*}
Q & \leftarrow Q_0 \\
\text{Unprocessed} & \leftarrow \mathcal{X} \\
\text{while } \text{Unprocessed} \neq \emptyset & \\
\text{choose } X_i \text{ from } \text{Unprocessed} & \\
Q_{\text{old}}(X_i) & \leftarrow Q(X_i) \\
\text{for } x_i \in \text{Val}(X_i) & \text{do} \\
Q(x_i) & \leftarrow \exp \left\{ \sum_{\phi \in \text{Sep}(\theta)} E_Q[\ln \phi_{x_i}] \right\} \\
\text{normalize } Q(X_i) \text{ to sum to one} & \\
\text{if } Q_{\text{old}}(X_i) \neq Q(X_i) & \text{then} \\
\text{Unprocessed} & \leftarrow \text{Unprocessed} \cup \{ \cup_{\phi \in \text{Sep}(\theta)} \text{Scope}(\phi) \} \\
\text{Unprocessed} & \leftarrow \text{Unprocessed} \setminus \{ X_i \} \\
\end{align*}

return \( Q \)

Mean Field Approximation

- Converges to one of typically many local minima.
- Easy to compute but sometimes is not good enough.
- It cannot describe complex posteriors (eg. XOR)
- We must use a richer class of distributions \( Q \).
Structured Mean Field

Exploiting Substructures

- If we use a distribution $Q$ that can capture some of the dependencies in $P$, we can get a better approximation.

Two possible substructures for $Q$

\[
Q(x) = \frac{1}{Z_Q} \prod_j \psi_j
\]

where $\psi_j$ is a factor with $\text{Scope}[\psi_j] = C_j$

and assume we have the set of potential scopes:

$\{C_j \subseteq \mathcal{X}: j = 1,\ldots,J\}$
Structured Mean Field

Exploiting Substructures

Given: \( Q(x) = \frac{1}{Z_Q} \prod_j \psi_j \)

And restriction: \( \sum_{c_j} \psi_j(c_j) = 1 \)

Then the potential \( \psi_j \) is locally optimal when:

\[
\psi_j(c_j) \propto \exp \left\{ E_Q \left[ \ln P_{\phi} \mid c_j \right] - \sum_{k \neq j} E_Q \left[ \ln \psi_k \mid c_j \right] \right\}
\]

Structured Mean Field

Exploiting Substructures

• Locality as Mean Fields:

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

where

\( A_j = \{ \phi \in F : Q \mid \neq (U_\phi \perp C_j) \} \)

and

\( B_j = \{ \psi_k : Q \mid \neq (C_k \perp C_j) \} \)
Structured Mean Field

Updating:

- Calculation of $Q(X_i)$ depends on clusters where $X_i$ belongs to. And on clusters overlapping $C_j$ (in $Q$). And on scopes $C_k$ dependent of $C_j$ (in $Q$ also).

![Diagram showing the relationship between clusters and scopes in a structured mean field model.]

Structured Mean Field

In other words, we want to compute $A_{1,1}$:

- $C_1 = \{A_{1,1}, A_{1,2}\}$
- $C_2 = \{A_{1,2}, A_{1,3}\}$
- $C_3 = \{A_{2,1}, A_{2,3}\}$

$A_j =$ Clusters $X_i$ belongs to (as standard mean field) i.e. $\{A_{1,1}, A_{1,2}\}$ and $\{A_{1,1}, A_{2,1}\}$

Clustering overlapping $C_j$ and those from PF. For example in this case $A_{1,2}$ in $C_j$ overlaps in $P$, thus we need to consider $\{A_{1,2}, A_{1,3}\}$ and $\{A_{1,1}, A_{2,3}\}$. The same occurs with $A_{1,3}$ and $A_{1,4}$.

$B_j =$ Clusters in $Q$ dependent on $C_j$. In this example every $C$ is independent from each other, therefore $B_j$ is empty.

![Diagram showing the relationship between clusters and scopes in a structured mean field model.]
Structured Mean Field

Again we want to compute $A_{1,b}$, assume the new substructure in $Q$:

Now we choose $C_1 = \{A_{1,1}, A_{1,2}, A_{1,3}, A_{1,4}, A_{2,1}\}$, $C_2 = \{A_{2,1}, A_{2,2}, A_{2,3}, A_{2,4}\}$

$A_j$: We consider the same clusters as before but now we add those overlapping with $A_{2,1}$, i.e. $\{A_{2,1}, A_{3,1}\}$ and $\{A_{2,1}, A_{2,2}\}$

$B_j$: Clusters in $Q$ dependent on $C_j$. Now we have $A_{2,1}$ (in $C_1$) overlapping with $A_{2,1}$ (in $C_2$). We need to subtract $A_{2,1}$ since we already used it in $A_j$.

Another example, we want to compute $a, b$:

Now we choose $C_1 = \{A, B\}$, $C_2 = \{C, D\}$

$A_j = \{\{A, B\}, \{A, D\}, \{B, C\}\}$

$B_j$: Empty, since $C_1$ and $C_2$ do not overlap.
Structured Mean Field

Exploiting Substructures

- Updates are relatively costly due to the consideration of structure.

Two approaches for updates:

- **Sequential**: Choose a factor and update it, then perform inference. It will converge.
- **Parallel**: Update all factors, then inference. It doesn’t guarantee convergence.

Structured Mean Field

Example:

\[
Q(A, B, C, D) = \frac{1}{Z_Q} \psi_1(A, B)\psi_2(C, D)
\]

\[
Q'(A, B, C, D) = \frac{1}{Z_Q} \phi_{ab}(A, B)\phi_{cd}(C, D)\psi'_1(A)\psi''_1(B)\psi'_2(C)\psi''_2(D)
\]

Structure (c) **cannot** be exploited (redundant)
Structured Mean Field

Refinement Theorem:
- Refines an initial approximating network by factorizing its factors into a product of factors and potentials from $P_F$.

- $\psi_k$ can be written as the product of two sets of factors:
  - Those in $P_F$ that are subsets of the scope of $\psi_k$.
  - Partially "covered" factors in $P_F$ by the scope of $\psi_j$ and other factors in $Q$.

Weigthed Mean Field

General Mixture Weights
- Idea:
  Instead of selecting one particular MF solution, we form a weighted average (a mixture) of several solutions.

- Enumerate the different MF-solutions by a hidden variable $a$, $Q(X|a)$.

- Assign mixture weights $Q(a)$.

\[ Q(X) = \sum_a Q(X | a)Q(a) \]
Weighthed Mean Field

Given \( Q(X) = \sum_a Q(X \mid a) Q(a) \)

under the constraint \( \sum_a Q(a) = 1 \)

Determine \( Q(a) \) such that \( D(Q\|P) \) is minimized:

\[
Q(a) \propto \exp \left[ - \log \frac{P(X \mid a)}{P(X)} \right]_{Q(X \mid a)}
\]

\[
\propto \exp \left[ - D(Q(X \mid a) \| P(X)) \right]
\]

Weighthed Mean Field

General Mixture Weights

• The previous formula means that different solutions \( Q(X \mid a) \) contribute to the global distribution \( Q(X) \) according to their distance to \( P(X) \).

• Note however, we are not optimizing \( Q(X \mid a) \) simultaneously.
**Weighthed Mean Field**

**Example: Noisy-OR**

- **Idea:**
  - Introducing auxiliary variational parameters that help in simplifying a complex objective function.

- **ln(x) ≤ λx - ln(λ) - 1**

|       | $X_1=1$ | $X_2=1$ | $X_3=1$ | $P(\cdot | X_2=1)$ |
|-------|---------|---------|---------|-----------------|
| $Q(\cdot | X_3=1, \alpha=1)$ | 0.137   | 0.973   | 0.555   | 0.528 |
| $Q(\cdot | X_3=1, \alpha=2)$ | 0.973   | 0.137   | 0.555   | 0.528 |

<table>
<thead>
<tr>
<th></th>
<th>$X_1=0$</th>
<th>$X_2=0$</th>
<th>$X_2=1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(X_1, X_2</td>
<td>X_3=1)$ compared to $Q(X_1, X_2</td>
<td>X_3=1)$</td>
<td>P: 0.005 Q: 0.023</td>
</tr>
<tr>
<td></td>
<td>$X_1=1$</td>
<td>$X_2=0$</td>
<td>$X_2=1$</td>
</tr>
<tr>
<td></td>
<td>P: 0.466 Q: 0.422</td>
<td>P: 0.063 Q: 0.138</td>
<td></td>
</tr>
</tbody>
</table>

**Variational Methods**

- **Idea:**
  - Introducing auxiliary variational parameters that help in simplifying a complex objective function.
Thank you!

Mean Field Approximation

Example from Wiegerinck:

Noisy-OR from Haft et al.:

<table>
<thead>
<tr>
<th></th>
<th>$X_1=1$</th>
<th></th>
<th>$X_2=1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MF-solution</td>
<td>0.137</td>
<td>exact</td>
<td>0.973</td>
</tr>
<tr>
<td>$Q(X=1)$</td>
<td></td>
<td>marginals</td>
<td>0.528</td>
</tr>
<tr>
<td>$P(X=1)$</td>
<td></td>
<td></td>
<td>0.528</td>
</tr>
</tbody>
</table>