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

Introduction

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 $Q(\mathrm{x})$ which is the best approximation of $P(\mathrm{x})$ and use $Q(\mathrm{x})$ to make inferences.

- Find $Q \in \mathrm{Q}$ that minimizes **F(P,Q)**

Mean Field Approximation

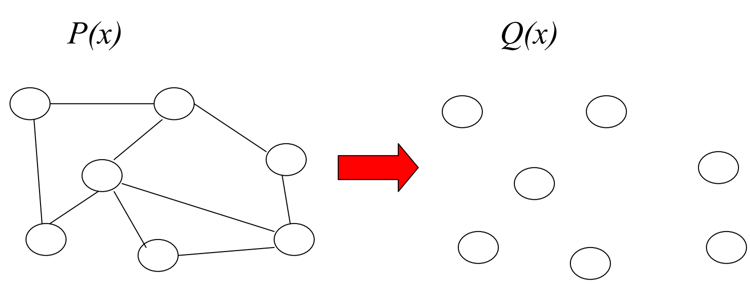
Assumptions:

- $Q(x)$ 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:

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

Mean Field Approximation

What happens when we apply mean field?



Mean Field Approximation

$$F(P,Q) = -\sum_{\phi \in F} \sum_{x_\phi} Q(x_\phi) \log \phi(x_\phi) + \sum_x Q(x) \log Q(x)$$

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

$$E(P,Q) = -\sum_{\phi \in F} \sum_{x_\phi} Q(x_\phi) \log \phi(x_\phi) = -\sum_{\phi \in F} \sum_{x_\phi} \left(\prod_{x_i \in x_\phi} Q(x_i) \right) \log \phi(x_\phi)$$

$$H(Q) = -\sum_x Q(x) \log Q(x) = -\sum_x \left(\prod_{i \in x} Q(x_i) \right) \log \left(\prod_{x_i \in x} Q(x_i) \right)$$

$$= -\sum_x \left(\prod_i Q(x_i) \right) \sum_i \log Q(x_i)$$

$$= -\sum_i \sum_{x_i} Q(x_i) \log Q(x_i)$$

$$= \sum_i H_{Q_i}(x_i)$$

Mean Field Approximation

$$F(P,Q) = -\sum_{\phi \in F} \sum_{x_\phi} Q(x_\phi) \log \phi(x_\phi) + \sum_x Q(x) \log Q(x)$$

$$E(P,Q) = -\sum_{\phi \in F} \sum_{x_\phi} \left(\prod_{x_i \in x_\phi} Q(x_i) \right) \log \phi\ (x_\phi)$$

$$H(Q) = -\sum_i \sum_{x_i} Q(x_i) \log Q(x_i)$$

Task: find $Q(x) = \prod_i Q_i(x_i)$ **minimizing F(P,Q)**

such that $\sum_{x_i} Q(x_i) = 1$

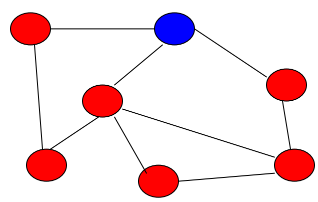
Solving: build a Lagrangian, differentiate and set to 0 !

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:

$$Q(x_i) = \frac{1}{Z_i} \exp\left\{ \sum_{\phi \in F} E_Q[\ln \phi \mid x_i] \right\} \qquad \text{MF-equation}$$

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



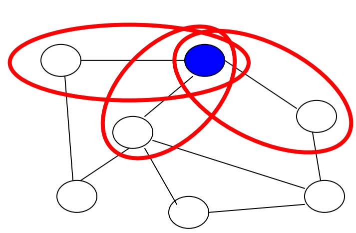
Mean Field Approximation

Locality:

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

$$Q(x_i) = \frac{1}{Z_i} \exp\left\{ \sum_{\phi : X_i \in Scope[\phi]} E_Q[\ln \phi(U_\phi, x_i)] \right\} \qquad \text{MF-equation simplified}$$

where $U_\phi = Scope\ [\phi]$



Calculation of $Q(X_i)$ depends only on clusters X_i belongs to

Mean Field Approximation

Solution: Iterate mean field equations

$$Q(x_i) = \frac{1}{Z_i} \exp\left\{ \sum_{\phi : X_i \in Scope[\phi]} E_Q\left[\ln \phi(U_\phi, x_i)\right] \right\}$$ MF-equation simplified

- 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

$$D(Q \mid P) = E_Q\left(\log \frac{Q(x)}{P(x)} \right)$$

$$D(Q \mid P) = E_Q\left(\log Q(x)\right) - E_Q(\log P(x))$$

$$D(Q \mid P) = E_Q\left(\log Q(\overline{X}_i)\right) - E_Q(\log P(\overline{X}_i)) + E_Q\left(\log Q(X_i)\right) - E_Q(\log P(X_i \mid \overline{X}_i))$$

Assume: $P(X) = P(X_i \mid \overline{X}_i) P(\overline{X}_i)$

$$D(Q \mid P) = E_{Q(\overline{X}_i)}\left(\log Q(\overline{X}_i)\right) - E_{Q(\overline{X}_i)}(\log P(\overline{X}_i)) + E_{Q(X_i)}\left(\log Q(X_i)\right) - E_{Q(X)}(\log P(X_i \mid \overline{X}_i))$$

Mean Field Approximation

Haft et al. paper:

- Optimize the KL divergence instead of the free energy

$$D(Q\,|\,P) = \boxed{E_{Q(\overline{X}_i)}\big(\log Q(\overline{X}_i)\big) - E_{Q(\overline{X}_i)}(\log P(\overline{X}_i))}$$
$$\boxed{+\,E_{Q(X_i)}\big(\log Q(X_i)\big) - E_{Q(X)}(\log P(X_i\,|\,\overline{X}_i))}$$

Does not depend on X_i

Depends on X_i

$$\frac{\partial D(Q\,|\,P)}{\partial Q(X_i)} = \frac{\partial}{\partial Q(X_i)} E_{Q(X_i)}\big(\log Q(X_i)\big) - E_{Q(X)}(\log P(X_i\,|\,\overline{X}_i))$$

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

Mean Field Approximation

Haft et al. paper:

$$Q(X_i) \propto \exp\Big(E_{Q(\overline{X}_i)}(\log P(X_i\,|\,\overline{X}_i))\Big) = \exp\big\langle \log P(X_i\,|\,\overline{X}_i)\big\rangle_{Q(\overline{X}_i)}$$

MF-equation

Locality:

$$Q(x_i) \propto \exp\langle \log P(x_i\,|\,M_i)\rangle_{Q(M_i)}$$

MF-equation simplified

where M is the Markov boundary.

Mean Field Approximation

Algorithm:

Procedure Mean-Field (
  $\mathcal{F}$, // factors that define $P_{\mathcal{F}}$
  Q_0 // Initial choice of Q
)
  $Q \leftarrow Q_0$
  $Unprocessed \leftarrow \mathcal{X}$
  while $Unprocessed \neq \emptyset$
    Choose X_i from $Unprocessed$
    $Q_{old}(X_i) \leftarrow Q(X_i)$
    for $x_i \in Val(X_i)$ **do**
      $Q(x_i) \leftarrow \exp\left\{\sum_{\phi: X_i \in Scope[\phi]} \boldsymbol{E}_Q[\ln \phi|_{x_i}]\right\}$
    Normalize $Q(X_i)$ to sum to one
    if $Q_{old}(X_i) \neq Q(X_i)$ **then**
      $Unprocessed \leftarrow Unprocessed \cup (\cup_{\phi: X_i \in Scope[\phi]} Scope[\phi])$
    $Unprocessed \leftarrow Unprocessed - \{X_i\}$
return Q

Mean Field Approximation

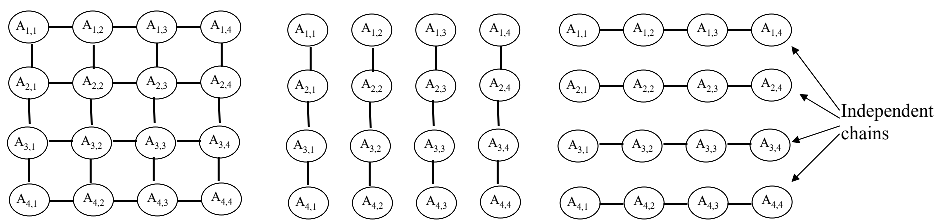
- Converges to one of typically many local minima.
- Easy to compute but sometimes is not good enough.
- It cannot describe complex posteriors (eg. XOR)
- We must use a richer class of distributions Q.

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



Distribution P Distribution $Q1$ Distribution $Q2$

Structured Mean Field

Exploiting Substructures

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

where ψ_j is a factor with $Scope[\psi_j] = C_j$.

and assume we have the set of potential scopes:

$\{C_j \subseteq \chi : j = 1, \ldots, J\}$

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 ψ_j is locally optimal when:

$$\psi_j(c_j) \propto \exp\left\{ E_Q\left[\ln P'_F \mid c_j\right] - \sum_{k \neq j} E_Q\left[\ln \psi_k \mid c_j\right] \right\}$$

Structured Mean Field

Exploiting Substructures

- Locality as Mean Fields:

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

where

$$A_j = \{\phi \in F : Q \not\models (U_\phi \perp C_j)\}$$

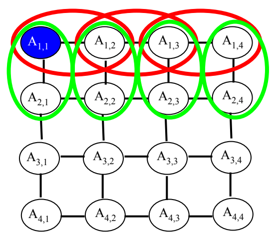
and

$$B_j = \{\psi_k : Q \not\models (C_k \perp C_j)\}$$

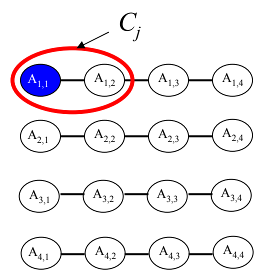
Structured Mean Field

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).



Distribution P



Distribution Q

Structured Mean Field

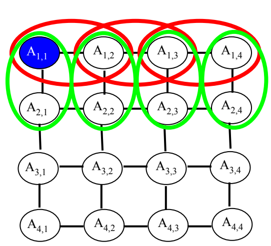
In other words, we want to compute $A_{1,1}$:

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

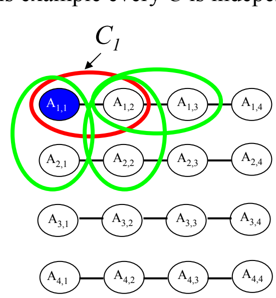
A_j = Clusters X_i belongs to (as standard mean field) i.e. $\{A_{1,1}\ A_{1,2}\}$ and $\{A_{1,1}\ A_{2,1}\}$

Clusters overlapping C_j and those from PF. For example in this case $A_{1,2}$ in C_1 overlaps in P_F, thus we need to consider $\{A_{1,2}\ A_{1,3}\}$ and $\{A_{1,1}\ A_{2,2}\}$. 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.



Distribution P

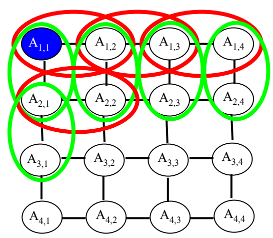


Distribution Q

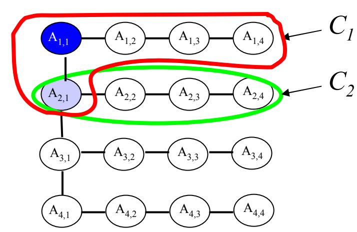
Structured Mean Field

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

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



Distribution P

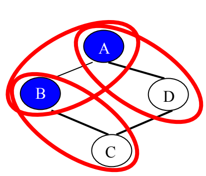


Distribution Q

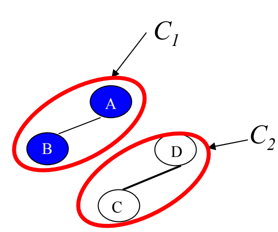
Structured Mean Field

Another example, we want to compute a,b:

- Now we choose $C_1 = \{A\ B\}$ $\quad C_2 = \{C\ D\}$
- $A_j = \{\ \{A\ B\}\ \{A\ D\}\ \{B\ C\}\}$
- B_j = Empty, since C_1 and C_2 do not overlap.



Distribution P



Distribution Q

Structured Mean Field

Exploiting Substructures

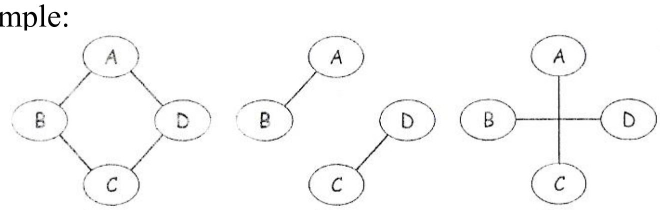
- Updates are relatively costly due to the consideration of structure.

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:



(a) $P_{\mathcal{F}}$ (b) Q_1 (c) Q_2

Structure (b) can be exploited:

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

$$Q'(A,B,C,D)=\frac{1}{Z_Q}\phi_{AB}(A,B)\phi_{CD}(C,D)\psi'_1(A)\psi''_1(B)\psi'_2(C)\psi''_2(D)$$

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.

- ψ_k can be written as the product of two sets of factors:
  - Those in P_F that are subsets of the scope of ψ_k.
  - Partially "covered" factors in P_F by the scope of ψ_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, $Q(\mathrm{X}|a)$.

- Assign mixture weights $Q(a)$.

$$Q(\mathrm{X}) = \sum_a Q(\mathrm{X} \mid a) Q(a)$$

Weigthed Mean Field

Given $Q(\mathrm{X}) = \sum_a Q(\mathrm{X} \mid a) Q(a)$

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

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

$$Q(a) \propto \exp\left[-\left\langle \log \frac{P(\mathrm{X} \mid a)}{P(\mathrm{X})} \right\rangle_{Q(\mathrm{X} \mid a)} \right]$$

$$\propto \exp\left[-D(Q(\mathrm{X} \mid a) \,\|\, P(\mathrm{X})) \right]$$

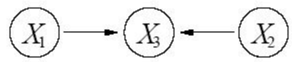
Weigthed Mean Field

General Mixture Weights

- The previous formula means that different solutions $Q(\mathrm{X}|a)$ contribute to the global distribution $Q(\mathrm{X})$ according to their distance to $P(\mathrm{X})$.
- Note however, we are **not** optimizing $Q(\mathrm{X}|a)$ simultaneously.

Weigthed Mean Field

Example: Noisy-OR



	first MF-solution $Q(\cdot\|X_3{=}1, a{=}1)$	second MF-solution $Q(\cdot\|X_3{=}1, a{=}2)$	marginals of MF-mixture $Q(\cdot\|X_3{=}1)$	exact marginals $P(\cdot\|X_3{=}1)$
$X_1{=}1$	0.137	0.973	0.555	0.528
$X_2{=}1$	0.973	0.137	0.555	0.528

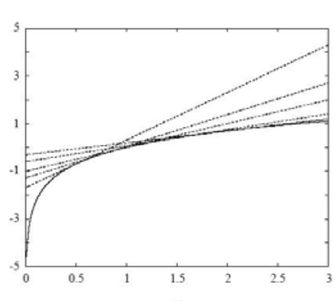
$P(X_1, X_2\|X_3 = 1)$ compared to $Q(X_1, X_2\|X_3 = 1)$	$X_2 = 0$	$X_2 = 1$
$X_1 = 0$	P: 0.005 Q: 0.023	P: 0.466 Q: 0.422
$X_1 = 1$	P: 0.466 Q: 0.422	P: 0.063 Q: 0.133

Variational Methods

Idea:

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

$$ln(x) \leq \lambda x - ln(\lambda) - 1$$

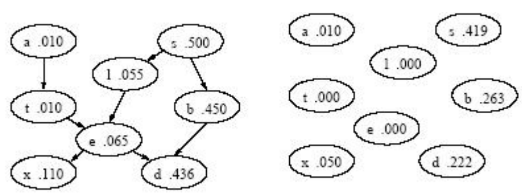


This upper bound allows to approximate *ln(x)* with a term that is linear in *x*.

Thank you!

Mean Field Approximation

Example from Wiegerinck:



(a) Target distribution (b) KL = 0.43

Noisy-OR from Haft et al.:

	first MF-solution $Q(\cdot\|X_3{=}1)$	exact marginals $P(\cdot\|X_3{=}1)$
$X_1{=}1$	0.137	0.528
$X_2{=}1$	0.973	0.528