Structured Variational Inference

Sargur Srihari
srihari@cedar.buffalo.edu

Topics

1. Structured Variational Approximations
    1. The Mean Field Approximation
        1. The Mean Field Energy
        2. Maximizing the energy functional: fixed point characterization
        3. Maximizing the energy functional: The Mean Field Algorithm
    2. Structured Approximations
        1. Fixed point characterization
        2. Optimization
        3. Simplifying the update equations
        4. Simplifying the family
        5. Selecting the approximation
    3. Local Variational Methods
        1. Variational bounds
        2. Variational variable elimination

Structured Variational Approximation

- Approximate inference methods based on belief propagation optimize an approximate energy functional over the class of pseudo marginals
  - But the pseudo marginals do not correspond to a globally coherent joint distribution Q
- The structured variational approach aims to optimize the energy functional over a family $\mathcal{Q}$ of coherent distributions Q
  - This family is chosen to be computationally tractable
    - Hence it is not sufficiently expressive to capture all of the information in P_Φ

Inference as Maximization

- We address the following maximization problem in Structured Variational Inference:

> Find Q ε $\mathcal{Q}$
>
> Maximizing $F[\tilde{P}_\Phi, Q]$
>
> where $\mathcal{Q}$ is a given family of distributions

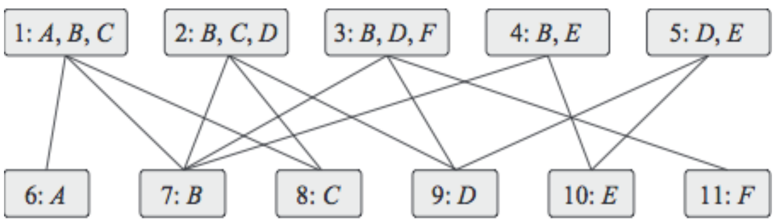
- We use the exact energy functional $F[P_\Phi, Q]$
    - which satisfies $\boldsymbol{D}(Q||P_\Phi) = \ln Z$ - $F[P_\Phi, Q]$
  - Thus maximizing the energy functional corresponds directly to obtaining a better approximation to P_Φ in terms of $\boldsymbol{D}(Q||P_\Phi)$

Parameter of Maximization

- Main parameter in maximization problem is choice of family Q
  - This choice introduces a trade-off
    - Families that are simpler, i.e., BNs and MNs of small tree width allow more efficient inference
    - If Q is too restrictive then it cannot represent distributions that are good approximations of P_Φ
      - Giving rise to a poor approximation Q
- Family is chosen to have enough structure
  - Hence called *structured variational approximation*
    - Variational calculus since we maximize over functions
    - Unlike belief propagation guaranteed to lower bound the log-partition function and guaranteed to converge

The Mean Field Approximation

- First approach considered is the mean field approximation
- It resembles the Bethe approximation to the mean field functional



Bethe bipartite graph; first layer of large clusters and second layer of univariate clusters

- The resulting algorithm performs message passing where the messages are over single variables
  - The form of the updates is different

Mean Field Class of Distributions

- The mean field algorithm finds the distribution Q which is closest to P_Φ in terms of $\boldsymbol{D}(Q||P_\Phi)$
- within the class of distributions representable as the product of independent marginals

$$Q(\chi) = \prod_i Q(X_i)$$

- Trade-off:
  - A fully factored distribution loses information
  - But we can easily evaluate any query
    - By a product of terms that involve variables in query
  - To represent Q we only need marginals of variables

Derivation of Mean Field Algorithm

- We consider the energy functional in the form

$$F\left[\tilde{P}_\Phi, Q\right] = E_Q\left[\ln \tilde{P}\left(\chi\right)\right] + H_Q\left(\chi\right) = \sum_{\phi \in \Phi} E_Q\left[\ln \phi\right] + H_Q\left(\chi\right)$$

  – where Q has the form of $Q(\chi) = \prod_i Q(X_i)$

A functional takes a function as input and produces a value as output, e.g., entropy

- We then characterize fixed points for each Q
- Thereby derive an iterative algorithm to find such fixed points
  – In fixed point iteration $x_{n+1} = f(x_n)$, $n = 0,1,2..$

Example of fixed pt: Newton's roots

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

Computing the Energy Functional

- The functional contains two terms: $F\left[\tilde{P}_{\Phi},Q\right]=\sum_{\phi\in\Phi}E_{Q}\left[\ln\phi\right]+H_{Q}\left(\chi\right)$
  1. The first is a sum of terms of the form $E_{U_\Phi \sim Q}[\ln \phi]$ where we need to evaluate

$$E_{U_\phi \sim Q}\left[\ln\phi\right]=\sum_{\boldsymbol{u}_\phi}Q(\boldsymbol{u}_\phi)\ln\phi(\boldsymbol{u}_\phi)$$
$$=\sum_{\boldsymbol{u}_\phi}\left(\prod_{Xi\in\boldsymbol{U}_\phi}Q(x_i)\right)\ln\phi(\boldsymbol{u}_\phi)$$

Notation: $P_{\Phi}\left(\chi\right)=\frac{1}{Z}\prod_{\phi\in\Phi}\phi\left(U_\phi\right)$

where $U_\phi = Scope\ (\phi)$

   - We can use the form of Q to compute $Q(u_\phi)$ as a product of marginals (performed in linear in no. of values of U_ϕ
  2. The term $H_Q(\chi)$ also decomposes in this case

If $Q(\chi)=\prod_i Q(X_i)$ then $H_Q(\chi)=\sum_i H_Q(X_i)$

Entropy definition:

$$H_Q(X_i)=E_Q\left[\ln\frac{1}{Q(X_i)}\right]$$

- Thus energy functional is a sum of expectations
  - Each one over a small set of variables

Complexity of Energy Functional

- Energy functional for a fully factored distribution Q can be written as a sum of expectations:

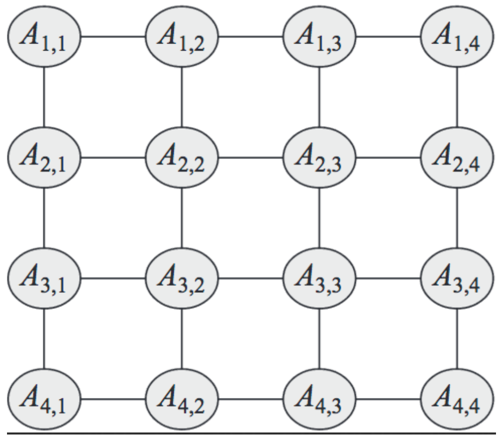
$$F\left[\tilde{P}_\Phi, Q\right] = \sum_{\boldsymbol{u}_\phi} \left(\prod_{Xi \in \boldsymbol{U}_\phi} Q(x_i) \right) \ln \phi(\boldsymbol{u}_\phi) + \sum_i E_Q\left[\ln \frac{1}{Q(X_i)} \right]$$

  - Each one over a small set of variables
  - Complexity depends on size of the factors in P_ϕ
    - Not on the topology of the network

$$P_\Phi(\chi) = \frac{1}{Z} \prod_{\phi \in \Phi} \phi(U_\phi)$$

- Thus the energy functional in this case can be represented/manipulated effectively
  - even when exact inference requires exponential time

Ex: Mean field energy for 4x4 grid



- Energy functional has the form

$$\boxed{\begin{aligned} F\left[\tilde{P}_{\Phi}, Q\right] = & \sum_{i \in \{1,2,3\}, j \in \{1,2,3,4\}} E_Q[\ln \phi(A_{i,j}, A_{i+1,j}) \\ & + \sum_{i \in \{1,2,3,4\}, j \in \{1,2,3\}} E_Q[\ln \phi(A_{i,j}, A_{i,j+1}) \\ & + \sum_{i \in \{1,2,3,4\}, j \in \{1,2,3,4\}} H_Q(A_{i,j}) \end{aligned}}$$

  - It involves only expectations over single variables and pairs of neighboring variables
    - The expression has the same form for an n x n grid
      - Thus although tree width for an n x n grid is exponential in n, the energy functional can be computed in $O(n^2)$, i.e., linear in no. of variables n^2

Maximizing Energy Functional: Fixed-point Characterization

- Task is to find distribution Q for which energy functional is maximized

Mean-Field

Find $Q \in \mathcal{Q}$

Maximizing $F[\tilde{P}_\Phi, Q]$

Subject to $Q(\chi) = \prod_i Q(X_i)$

$$\sum_{x_i} Q(x_i) = 1 \quad \forall i$$

  – We use Lagrange multipliers to characterize stationary points of $F[\tilde{P}_\Phi, Q]$
    - The structure of Q allows us to consider the optimal value of each component given the rest

Mean Field Theorem: maximum of $Q(X_i)$

- *Thm*: $Q(X_i)$ is a local max of Mean-Field given $\{Q(X_j)\}_{j \neq i}$ iff

$$Q(x_i) = \frac{1}{Z_i} \exp\left\{ \sum_{\phi \in \Phi} E_{\chi \sim \Phi}\left[\ln \phi \mid x_i \right] \right\}$$

where Z_i is a local normalizing constant and $E_{\chi \sim Q}[\ln \boldsymbol{\phi} | x_i]$ is the conditional expectation given value x_i

$$E_{\chi \sim Q}[\ln \phi \mid x_i] = \sum_{\boldsymbol{u}_\phi} Q(\boldsymbol{u}_\phi \mid x_i) \ln(\boldsymbol{u}_\phi)$$

- Proof:

1. Objective function is: $F\left[\tilde{P}_\Phi, Q\right] = \sum_{\phi \in \Phi} E_Q\left[\ln \phi\right] + H_Q\left(\chi\right)$
2. Restricting attention to $Q(X_i)$ terms: $F_i\left[Q\right] = \sum_{\phi \in \Phi} E_{U_\phi \sim Q}\left[\ln \phi\right] + H_Q\left(X_i\right)$
3. To optimize $Q(X_i)$ define Lagrangian of terms in $F[\tilde{P}_\Phi, Q]$

$$L_i\left[Q\right] = \sum_{\phi \in \Phi} E_{U_\phi \sim Q}\left[\ln \phi\right] + H_Q(X_i) + \lambda\left(Q(x_i) - 1\right)$$

4. For derivatives use lemma: If $Q(\chi) = \prod_i Q(X_i)$ then for any function f with scope U: $\dfrac{\partial}{\partial Q(x_i)} E_{U \sim Q}\left[f(U)\right] = E_{U \sim Q}\left[f(U) \mid x_i\right]$
5. Using lemma & standard derivatives of entropy:

$$\frac{\partial}{\partial Q(x_i)} L_i = \sum_{\phi \in \Phi} E_{\chi \sim Q}\left[\ln \phi \mid x_i\right] - \ln Q(x_i) - 1 + \lambda$$

6. Set derivatives to 0 & rearrange: $\ln Q(x_i) = \lambda - 1 + \sum_{\phi \in \Phi} E_{\chi \sim Q}\left[\ln \phi \mid x_i\right]$
7. Take exponents of both sides, and λ constant.
8. Solution: $Q(X_i)$ is maximum since $\Sigma E_{U\varphi}$ is linear & H_Q is concave ■

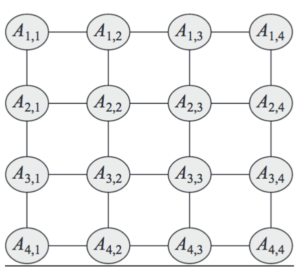
Corollary of Mean-Field Theorem

- To convert $Q(x_i)=\frac{1}{Z_i}\exp\left\{\sum_{\phi\in\Phi} E_{\chi\sim\Phi}\left[\ln\phi \mid x_i\right]\right\}$ into update algorithm
  - Observe that if $X_i \notin \text{Scope}(\phi)$ then $E_{U_\phi\sim Q}\left[\ln\phi \mid x_i\right]=E_{U_\phi\sim Q}\left[\ln\phi_i\right]$
    - i.e., expectation of such factors are independent of x_i

Cor. In mean field approx. $Q(X_i)$ is a local maximum iff

$$Q(x_i)=\frac{1}{Z_i}\exp\left\{\sum_{\phi:X_i\in Scope\{\phi\}} E_{(U_\phi-\{X_i\})\sim\Phi}\left[\ln\phi(U_\phi,x_i)\right]\right\}$$

- This representation shows that $Q(X_i)$ have to be consistent with expectation of the potentials in which it appears
  - In grid network, it implies that $Q(A_{i,j})$ is the product of four terms



$$Q(a_{i,j})=\frac{1}{Z_{i,j}}\exp\left\{\begin{array}{l}\sum_{a_{i-1,j}} Q(a_{i-1,j})\ln\left(\phi(a_{i-1,j},a_{i,j})\right)+\\ \sum_{a_{i,j-1}} Q(a_{i,j-1})\ln\left(\phi(a_{i,j-1},a_{i,j})\right)+\\ \sum_{a_{i+1,j}} Q(a_{i+1,j})\ln\left(\phi(a_{i+1,j},a_{i,j})\right)+\\ \sum_{a_{i,j+1}} Q(a_{i,j+1})\ln\left(\phi(a_{i,j},a_{i,j+1})\right)\end{array}\right\}$$

Each term is a geometric average of one of the potentials involving $A_{i,j}$

From Mean Field to Update Algorithm

- We now have tools for algorithm to find max $\boxed{F[\tilde{P}_\Phi, Q]}$
  - They are: $\boxed{Q(x_i) = \frac{1}{Z_i}\exp\left\{\sum_{\phi: X_i \in Scope\{\phi\}} E_{(U_\phi - \{X_i\}) \sim \Phi}\left[\ln \phi(U_\phi, x_i)\right]\right\}}$ and $\boxed{Q(a_{i,j}) = \frac{1}{Z_{i,j}}\exp\left\{\begin{array}{l} \sum_{a_{i-1,j}} Q(a_{i-1,j}) \ln\left(\phi(a_{i-1,j}, a_{i,j})\right) + \\ \sum_{a_{i,j-1}} Q(a_{i,j-1}) \ln\left(\phi(a_{i,j-1}, a_{i,j})\right) + \\ \sum_{a_{i+1,j}} Q(a_{i+1,j}) \ln\left(\phi(a_{i+1,j}, a_{i,j})\right) + \\ \sum_{a_{i,j+1}} Q(a_{i,j+1}) \ln\left(\phi(a_{i,j}, a_{i,j+1})\right) \end{array}\right\}}$
  - Term within exponential
    - is easily evaluated by
      - interactions between neighbors of $A_{i,j}$ and values they can take
    - RHS has expectations of variables not involving X_i
  - Resulting $Q(X_i)$ is optimal given all other values
    - Simply evaluate exponent terms for each value of x_i, normalize results to sum to 1 and assign them to $Q(X_i)$
    - Consequently we reach optimal $Q(X_i)$ in one easy step
  - To optimize relative to all variables, embed step in iterated coordinate ascent algorithm
      - Optimize single marginal at a time, given fixed choices of others

Algorithm: Mean-Field Approximation

- **Procedure** Mean-Field {Φ, //factors that define P_Φ,
  Q_0 //initial choice of Q }

1. $Q \leftarrow Q_0$

2. Unprocessed $\leftarrow \chi$

3. while *Unprocessed* $\neq \emptyset$

4. Choose X_i from *Unprocessed*

5. $Q_{\text{old}}(X_i) \leftarrow Q(X_i)$

6. for $x_i \in Val(X_i)$ **do**

7. $Q(x_i) \leftarrow$ $\boxed{\exp\left\{ \sum_{\phi: X_i \in Scope\{\phi\}} E_{(U_\phi - \{X_i\}) \sim \Phi}\left[\ln \phi(U_\phi, x_i)\right]\right\}}$

8. Normalize $Q(X_i)$ to sum to one

9. if $Q_{\text{old}}(X_i) \neq Q(X_i)$ **then**

10. Unprocessed $\leftarrow$ *Unprocessed* $\cup \left(\cup_{\phi: X_i \in Scope[\phi]} Scope[\phi]\right)$

11. Unprocessed $\leftarrow$ *Unprocessed*-$\{X_i\}$

- **return** Q

Observations about Algorithm

- A single optimization of $Q(X_i)$ does not suffice.
  - A subsequent modification of another marginal $Q(X_i)$ may result in a different optimal parameterization for $Q(X_i)$
  - Thus algorithm repeats the steps until convergence
- Each iteration of Mean-Field results in a better approximation Q to the target density P_Φ guaranteeing convergence
- The convergence points are local maxima
  - Not necessarily global

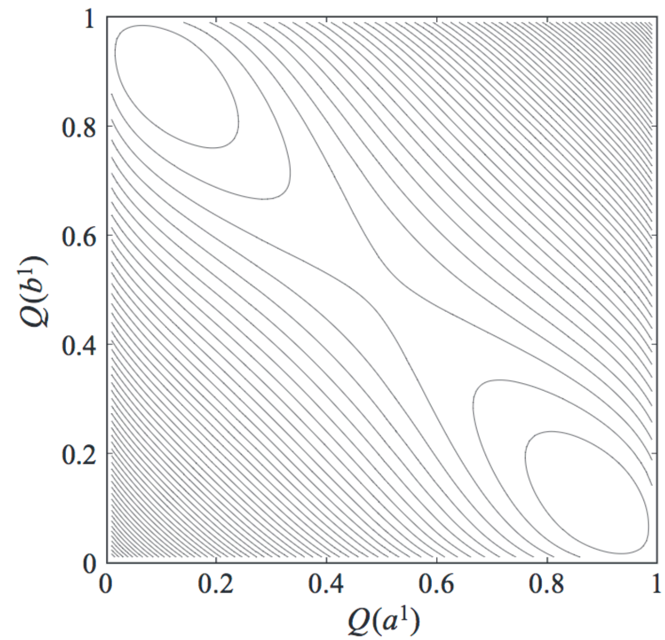
Ex: Mean-Field Energy Functional

- A distribution P_Φ

  $P(a,b) = 0.5 - \varepsilon$ if $a \neq b$

  $P(a,b) = \varepsilon$ if $a = b$

  – Which is a Noisy XOR



- Mean field Energy $F[\tilde{P}_\Phi, Q]$
- As XOR, P_Φ cannot be approximated by a product of marginals
- But energy potential surface has two peaks at $a \neq b$

Structured Approximations

- Although Mean Field Algorithm provides an easy approximation method, we are forcing Q to be a very simple distribution
  - All variables being independent of each other in Q leads to very poor approximations
- If we use a Q that can capture some dependencies in P_Φ we can get a better approximation
  - Thus explore approximations inbetween mean field and exact inference
  - Using network structures of different complexity

Fixed-point characterization

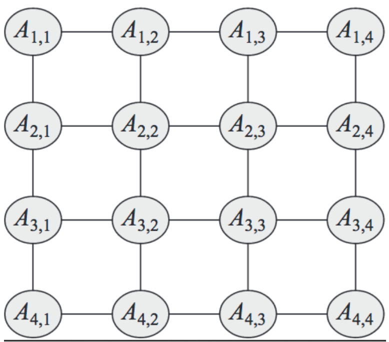
- Focus on using MNs instead of BNs
- Parameterized Gibbs distributions
  - Not restricted to factors over maximal cliques
- Form of variational approximation
  - Q is from a Gibbs parametric family
  - We are given a set of potential scopes $\{C_j \subseteq \chi : j = 1, \ldots, J\}$
  - We choose an approximation Q that has the form

$$Q(\chi) = \frac{1}{Z_Q} \prod_{j=1}^{J} \Psi_j$$

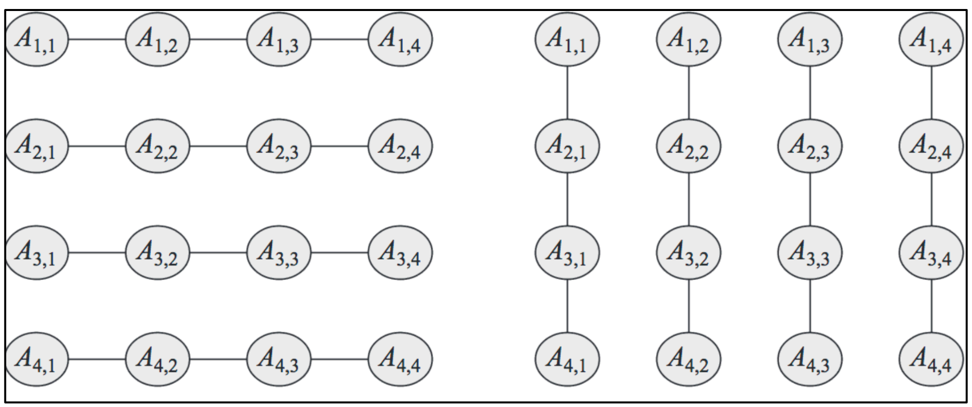
    - where Ψ_j is a factor with scope $Scope[\Psi_j] = C_j$

Ex: Grid Network

- Given:



  - There are many possible approximate network structures, two of them are:



Both a collection of independent chain structures

Inference is linear

Not much worse than mean field

Method of Structured Variational

- Decide on form of potentials ϕ for family Q
- We consider energy functional $F\left[\tilde{P}_{\Phi}, Q\right] = \sum_{\phi \in \Phi} E_Q\left[\ln \phi\right] + H_Q\left(\chi\right)$ for a distribution Q in this family
  - Characterize the stationary points of functional
  - Use those to derive iterative optimization algorithm
- Evaluating terms that involve $E_{U_\phi \sim Q}\left[\ln \phi\right]$ requires performing expectations wrt variables in $Scope[\phi]$
  - Complexity of expectation depends on structure of approximating distribution
    - Assume we can solve this problem using exact inference