CS 3750 Advanced Topics in ML
Lecture 7

Approximations as optimizations

Milos Hauskrecht
milos@cs.pitt.edu
5329 Sennott Square

Approximation as an optimization

Idea:

- Assume we have the distribution P(x) that can be represented compactly, but inference is not efficient
- Assume we have a class of distributions Q(x) that are easier to work with in inferences
- **Objective:** We want to find Q(x) which is the best approximation of P(x) and use Q(x) to make inferences

- How to define what is the best Q(x) ?
- The distance in between two distributions is measured by a **relative entropy** (**Kullback-Leibler divergence**)

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

KL divergence

KL divergence:

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

$$\sum_x Q(x) \log \frac{Q(x)}{P(x)} =$$

$$\sum_x Q(x) \log Q(x) - \sum_x Q(x) \log P(x)$$

Asymmetric measure:

$$D(P \mid Q) \neq D(Q \mid P)$$

Helmholz Free energy

Assume: $P(x) = \frac{1}{Z} \prod_{c \in cliques} \phi_c(x_c)$

$Q(x)$

$$D(Q \mid P) = \sum_x Q(x) \log Q(x) - \sum_x Q(x) \log P(x)$$

$$D(Q \mid P) = \sum_x Q(x) \log Q(x) - \sum_x Q(x) \log\left(\frac{1}{Z} \prod_{c \in cliques} \phi_c(x_c)\right)$$

$$D(Q \mid P) = (\log Z) \sum_x Q(x) - \sum_x Q(x)\left(\sum_{c \in cliques} \log \phi_c(x_c)\right) + \sum_x Q(x) \log Q(x)$$

$$D(Q \mid P) = +\log(Z) - \sum_x Q(x)\left(\sum_{c \in cliques} \log \phi_c(x_c)\right) + \sum_x Q(x) \log Q(x)$$

Helmholz free energy

Assume: $P(x) = \frac{1}{Z} \prod_{c \in cliques} \phi_c(x_c)$ and $Q(x)$

$$D(Q \mid P) = +\log(Z) - \sum_x Q(x) \left(\sum_{c \in cliques} \log \phi_c(x_x) \right) + \sum_x Q(x) \log Q(x)$$

$$\log Z = D(Q \mid P) + \sum_x Q(x) \left(\sum_{c \in cliques} \log \phi_c(x_c) \right) - \sum_x Q(x) \log Q(x)$$

$$\log Z = D(Q \mid P) + \underbrace{\sum_{c \in cliques} \sum_{x_c} Q(x_c) \log \phi_c(x_c) - \sum_x Q(x) \log Q(x)}_{\text{Helmholz free energy} \quad -F(P,Q)}$$

Helmholz free energy

Assume: $P(x) = \frac{1}{Z} \prod_{c \in cliques} \phi_c(x_c)$ and $Q(x)$

$$D(Q \mid P) = +\log(Z) - \sum_x Q(x) \left(\sum_{c \in cliques} \log \phi_c(x_x) \right) + \sum_x Q(x) \log Q(x)$$

$$\log Z = D(Q \mid P) + \sum_x Q(x) \left(\sum_{c \in cliques} \log \phi_c(x_c) \right) - \sum_x Q(x) \log Q(x)$$

$$\log Z = D(Q \mid P) + \sum_{c \in cliques} \sum_{x_c} Q(x_c) \log \phi_c(x_c) - \sum_x Q(x) \log Q(x)$$

Helmholz free energy

$$F(P,Q) = -\sum_{c \in cliques} \sum_{x_c} Q(x_c) \log \phi_c(x_c) + \sum_x Q(x) \log Q(x)$$

$$F(P,Q) = \underbrace{E(P,Q)}_{\text{energy}} - \underbrace{H_Q(Q)}_{\text{entropy}}$$

Helmholz free energy

Assume: $P(x) = \frac{1}{Z} \prod_{c \in cliques} \phi_c(x_c)$ and $Q(x)$

$$\log Z = D(Q \mid P) + \sum_{c \in cliques} \sum_{x_c} Q(x_c) \log \phi_c(x_c) - \sum_{x} Q(x) \log Q(x)$$

$$\log Z = D(Q \mid P) - E(P,Q) + H_Q(Q)$$

$$\underbrace{\log Z}_{} = \underbrace{D(Q \mid P)}_{} - \underbrace{F(P,Q)}_{}$$

Constant wrt Q = KL divergence + Helmholz free energy

Increase the distance → increase the HF energy
Decrease the distance → decrease the HF energy

Helmholz free energy

Assume: $P(x) = \frac{1}{Z} \prod_{c \in cliques} \phi_c(x_c)$ and $Q(x)$

$$\underbrace{\log Z}_{} = \underbrace{D(Q \mid P)}_{} - \underbrace{F(P,Q)}_{}$$

Constant wrt Q = KL divergence - Helmholz free energy

Increase the distance → increase the HF energy
Decrease the distance → decrease the HF energy

The problem of finding the best Q(x) approximating P(x) can be cast as the problem of minimizing the HF energy F(Q,P)

- **Note that -F(P,Q) is a lower bound of log Z**

$$\log Z \geq -F(P,Q)$$

Optimization of free energy

Approximations:

- **we want to find Q(x) optimizing F(P,Q)**

$$F(P,Q) = -\sum_x Q(x)\left(\sum_{c \in cliques} \log \phi_c(x_c)\right) + \sum_x Q(x)\log Q(x)$$

$$F(P,Q) = -\sum_{c \in cliques} \sum_{x_c} Q(x_c) \log \phi_c(x_c) + \sum_x Q(x)\log Q(x)$$

- **Solution: variational approximations**
  - **Tractable approximations of the Helmholz free energy**
- **Two solution methods:**
  1. **Approximate** $F(P,Q)$ with $\widetilde{F}(P,Q)$
  2. **Choose a simpler form of Q that is easy to work with**

Kikuchi approximation

Goal: Approximate $F(P,Q)$ with $\widetilde{F}(P,Q)$

$$F(P,Q) = -\sum_{c \in cliques} \sum_{x_c} Q(x_c) \log \phi_c(x_c) + \sum_x Q(x)\log Q(x)$$

$$F(P,Q) = E(P,Q) - H_Q(Q)$$

$$H_Q(Q) = -\sum_x Q(x)\log Q(x)$$

- **Kikuchi approximation:**

$$F_{kik}(P,Q) = E(P,Q) - H_{kik}(Q)$$

$$H_{kik}(Q) = -\sum_{c \in cliques} \sum_{x_c} Q(x_c)\log Q(x_c) - \sum_{\xi \in overlaps} \sum_{x_\xi} u_\xi Q(x_\xi)\log Q(x_\xi)$$

Sum over cliques + corrections for overlaps

where $Q(x_c)$ and $Q(x_\xi)$ **are properly normalized**

and u_ξ - Moebius (overcounting) number $= 1 - \sum_{\xi \subset \xi'} u_{\xi'} \quad \forall \xi$

Kikuchi approximation

- **Kikuchi approximation:**

$$F_{kik}(P,Q) = E(P,Q) - H_{kik}(Q)$$

$$H_{kik}(Q) = -\sum_{c \in cliques} \sum_{x_c} Q(x_c) \log Q(x_c) - \sum_{\xi \in overlaps} \sum_{x_\xi} u_\xi Q(x_\xi) \log Q(x_\xi)$$

Sum over cliques + corrections for overlaps

where $Q(x_c)$ and $Q(x_\xi)$ **are properly normalized**

and u_ξ - Moebius (overcounting) number = $1 - \sum_{\xi \subset \xi'} u_{\xi'} \quad \forall \xi$

- **Normalization conditions:**
  - **For all subsets γ' ⊂ γ it holds:** $\sum_{x_{\gamma \setminus \gamma'}} Q_\gamma(x_\gamma) = Q_{\gamma'}(x_{\gamma'})$
  - **For all clusters γ it holds:** $\sum_{x_\gamma} Q_\gamma(x_\gamma) = 1$

Kikuchi approximation

- **Kikuchi approximation:**

$$F_{kik}(P,Q) = E(P,Q) - H_{kik}(Q)$$

$$H_{kik}(Q) = -\sum_{c \in cliques} \sum_{x_c} Q(x_c) \log Q(x_c) - \sum_{\xi \in overlaps} \sum_{x_\xi} u_\xi Q(x_\xi) \log Q(x_\xi)$$

Sum over cliques + corrections for overlaps

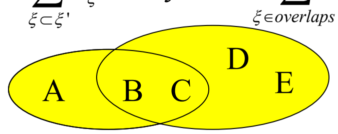
where $Q(x_c)$ and $Q(x_\xi)$ **are properly normalized**

and u_ξ - Moebius number

- **Moebius (overcounting) number:**

$u_\xi = 1$ if $\xi = c$ $\qquad - \sum_{c \in cliques} \sum_{x_c} 1\, Q(x_c) \log Q(x_c)$

$u_\xi = 1 - \sum_{\xi \subset \xi'} u_{\xi'} \quad \forall \xi$ $\qquad - \sum_{\xi \in overlaps} \sum_{x_\xi} u_\xi Q(x_\xi) \log Q(x_\xi)$



Bethe approximation

Goal: Approximate $F(P,Q)$ with $\widetilde{F}(P,Q)$

- **Kikuchi approximation:**

$$F_{kik}(P,Q) = E(P,Q) - H_{kik}(Q)$$

$$H_{kik}(Q) = -\sum_{c \in cliques} \sum_{x_c} Q(x_c) \log Q(x_c) - \sum_{\xi \in overlaps} \sum_{x_\xi} u_\xi Q(x_\xi) \log Q(x_\xi)$$

Sum over cliques + corrections for overlaps

u_ξ - Moebius number = $1 - \sum_{\xi \subset \xi'} u_{\xi'} \quad \forall \xi$

- **Bethe approximation (overlaps restricted to disjunct subsets, typically singletons)**

$$F_{Bethe}(P,Q) = E(P,Q) - H_{Bethe}(Q)$$

$$H_{Bethe}(Q) = -\sum_{c \in cliques} \sum_{x_c} Q(x_c) \log Q(x_c) + \sum_{\xi} \sum_{x_\xi} Q(x_\xi) \log Q(x_\xi)$$

Sum over cliques + corrections over disjunct subsets

Optimization of free energy

Approximations:

- **we want to find Q(x) optimizing F(P,Q)**

$$F(P,Q) = -\sum_{x} Q(x) \left(\sum_{c \in cliques} \log \phi_c(x_c) \right) + \sum_{x} Q(x) \log Q(x)$$

$$F(P,Q) = -\sum_{c \in cliques} \sum_{x_c} Q(x_c) \log \phi_c(x_c) + \sum_{x} Q(x) \log Q(x)$$

Solution 2:

Choose a simpler form of Q that is easy to work with

- **Example: a mean field approximation**

$$Q(x) = \prod_{i} Q_i(x_i)$$

If x_i **is binary than** $Q(x_i = 1) = \lambda_i$ **and** $Q(x_i = 0) = 1 - \lambda_i$

Mean Field approximation

$$F(P,Q) = -\sum_{c \in cliques} \sum_{x_c} Q(x_c) \log \phi_c(x_c) + \sum_x Q(x) \log Q(x)$$

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

$$E(P,Q) = -\sum_{c \in cliques} \sum_{x_c} Q(x_c) \log \phi_c(x_c) = -\sum_{c \in cliques} \sum_{x_c} \left(\prod_{x_i \in x_c} Q(x_i) \right) \log \phi_c(x_c)$$

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

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

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

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

Mean Field approximation

$$F(P,Q) = -\sum_{c \in cliques} \sum_{x_c} Q(x_c) \log \phi_c(x_c) + \sum_x Q(x) \log Q(x)$$

$$E(P,Q) = -\sum_{c \in cliques} \sum_{x_c} \left(\prod_{x_i \in x_c} Q(x_i) \right) \log \phi_c(x_c)$$

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

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

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

Solving: build a Lagrangian, differentiate and set to 0 !
- Leads to mean field equations that are iteratively solved

Mean Field approximation

$$F(P,Q) = -\sum_{c \in cliques} \sum_{x_c} Q(x_c) \log \phi_c(x_c) + \sum_x Q(x) \log Q(x)$$

$$E(P,Q) = -\sum_{c \in cliques} \sum_{x_c} \left(\prod_{x_i \in x_c} Q(x_i) \right) \log \phi_c(x_c)$$

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

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

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

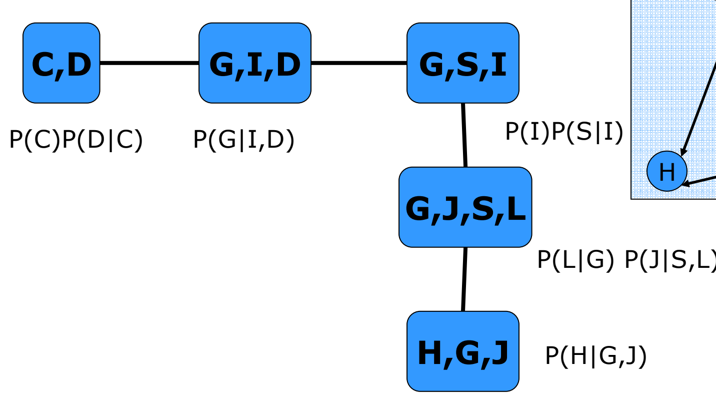
Solving: build a Lagrangian, differentiate set to 0 !

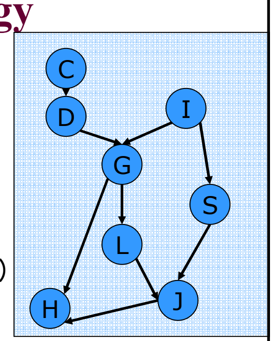
- Leads to mean field equations that are solved iteratively

Optimization of free energy

Example of an exact inference:

- **Assume a cluster tree**



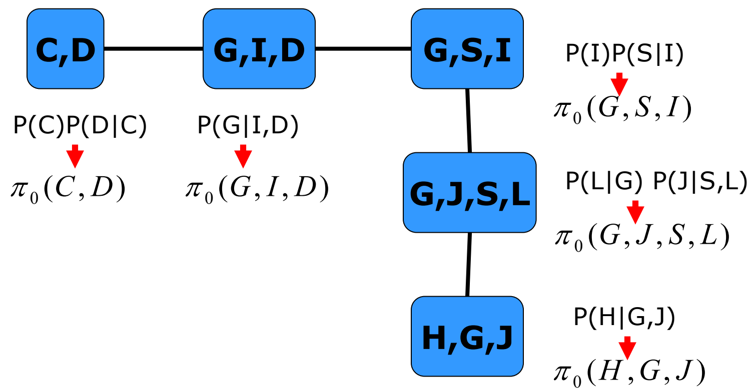


$$F(P,Q) = -\sum_{c \in cliques} \sum_{x_c} Q(x_c) \log \phi_c(x_c) + \sum_x Q(x) \log Q(x)$$

Optimization of free energy

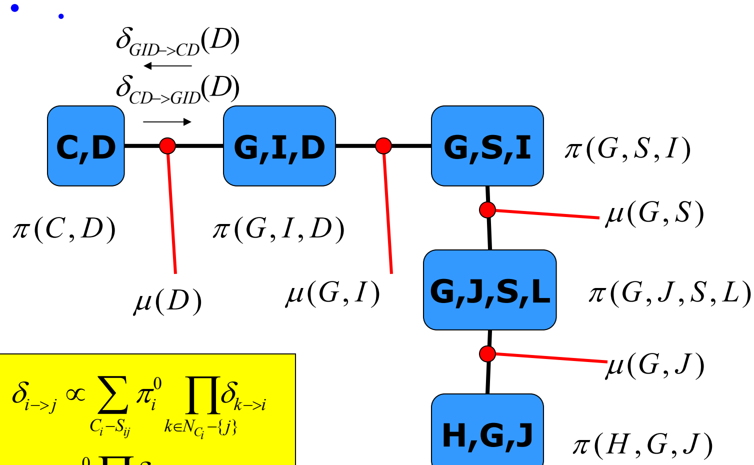
Example of an exact inference:

- **Assume a cluster tree**



Belief propagation: messages and potentials

- .



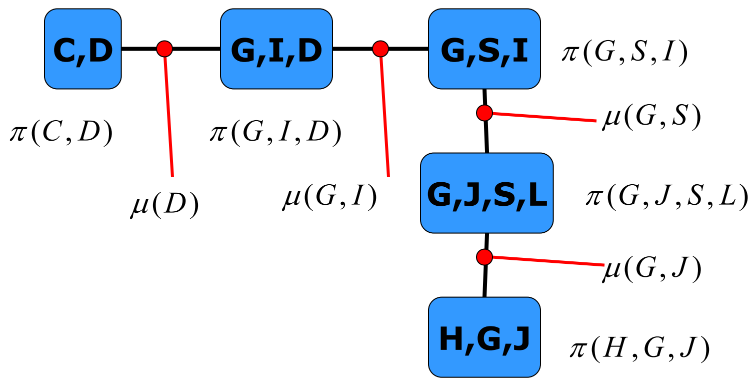
$$\delta_{i->j} \propto \sum_{C_i - S_{ij}} \pi_i^0 \prod_{k \in N_{C_i} - \{j\}} \delta_{k->i}$$

$$\pi_i \propto \pi_i^0 \prod_{k \in N_{C_i}} \delta_{k->i}$$

$$\mu_{ij} = \delta_{j->i} \delta_{i->j}$$

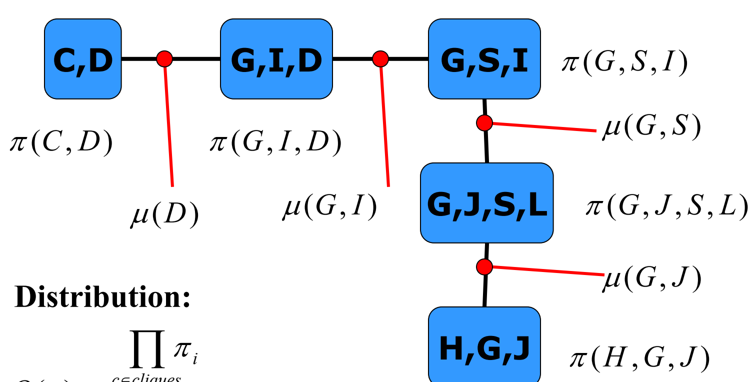
Belief propagation

- **Assume a cluster tree and potentials at the end of the belief propagation**



Optimization of free energy

- **Assume a cluster tree and potentials at the end of the belief propagation**

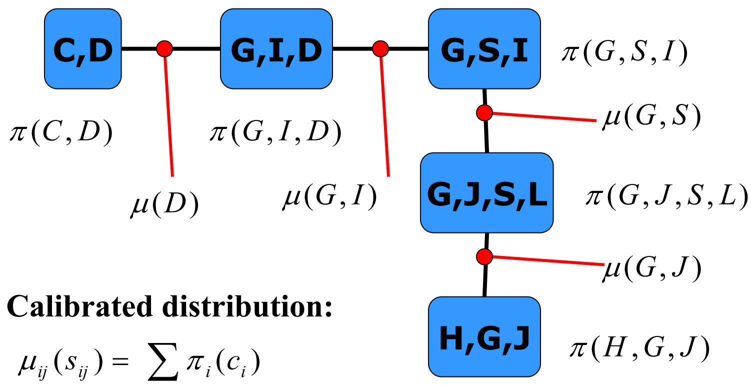


Distribution:

$$Q(x) = \frac{\prod_{c \in cliques} \pi_i}{\prod_{C_i -- C_j} \mu_{ij}}$$

Optimization of free energy

- **Assume a cluster tree and potentials at the end of the belief propagation**



Calibrated distribution:

$$\mu_{ij}(s_{ij}) = \sum_{C_i - S_{ij}} \pi_i(c_i)$$

for all clusters containing: S_{ij}

Optimization of free energy

Example of an exact inference:

- **Assume a Kikuchi approximation**

$$\widetilde{F}(P,Q) = -\sum_{c \in cliques} \sum_{x_c} Q(x_c) \log \phi_c(x_c) +$$

$$+ \sum_{c \in cliques} \sum_{x_c} Q(x_c) \log Q(x_c) + \sum_{\xi \in overlaps} \sum_{x_\xi} u_\xi Q(x_\xi) \log Q(x_\xi)$$

$$\widetilde{F}(P,Q) = -\sum_{c \in cliques} \sum_{x_c} \pi_c(x_c) \log \pi_c^0(x_c) +$$

$$+ \sum_{c \in cliques} \sum_{x_c} \pi_c(x_c) \log \pi_c(x_c) - \sum_{s \in overlaps} \sum_{x_s} \mu_s(x_s) \log \mu_s(x_s)$$

$$\widetilde{F}(P,Q) = F(P,Q)$$

Why ? Substitute Q(x) to the F(P,Q) !!

Optimization of free energy

Example of an exact inference:

- **Assume a Kikuchi approximation**

$$\tilde{F}(P,Q) = -\sum_{c \in cliques} \sum_{x_c} Q(x_c) \log \phi_c(x_c) +$$

$$+ \sum_{c \in cliques} \sum_{x_c} Q(x_c) \log Q(x_c) + \ldots$$

Note: Overlaps on disjoint subsets are equal to the Bethe approximation

$$\tilde{F}(P,Q) = -\sum_{c \in cliques} \sum_{x_c} \pi_c(x_c) \log \pi_c^0(x_c) +$$

$$+ \sum_{c \in cliques} \sum_{x_c} \pi_c(x_c) \log \pi_c(x_c) - \sum_{s \in overlaps} \sum_{x_s} \mu_s(x_s) \log \mu_s(x_s)$$

$$\tilde{F}(P,Q) = F(P,Q)$$

Why ? Substitute Q(x) to the F(P,Q) !!

Optimization of free energy

Example of an exact inference:

$$\tilde{F}(P,Q) = -\sum_{c \in cliques} \sum_{x_c} \pi_c(x_c) \log \pi_c^0(x_c) +$$

$$+ \sum_{c \in cliques} \sum_{x_c} \pi_c(x_c) \log \pi_c(x_c) - \sum_{s \in overlaps} \sum_{x_s} \mu_s(x_s) \log \mu_s(x_s)$$

$$\tilde{F}(P,Q) = F(P,Q)$$

Why ? Substitute Q(x) to the F(P,Q) !!

$$Q(x) = \frac{\prod_{c \in cliques} \pi_i}{\prod_{C_i - -C_j} \mu_{ij}}$$

$$F(P,Q) = -\sum_{c \in cliques} \sum_{x_c} Q(x_c) \log \phi_c(x_c) + \sum_{x} Q(x) \log Q(x)$$

$$F(P,Q) = -\sum_{c \in cliques} \sum_{x_c} \pi_c(x_c) \log \phi_c(x_c) +$$

$$+ \sum_{c \in cliques} \sum_{x_c} \pi_c(x_c) \log \pi_c(x_c) - \sum_{s \in S_i - S_j} \sum_{x_c} \mu_s(x_s) \log \mu_s(x_s)$$

Optimization of free energy

Optimization:

Find Q
Minimizing $\tilde{F}(P,Q)$

Subject to
$$\mu_s = \sum_{c_i - s} \pi_i \quad \text{for all } s$$
$$\sum_{x_c} \pi_c(x_c) = 1 \quad \text{for all } c$$

Assures calibration