Probabilistic Graphical Models (Cmput 651): Global Approximate Inference

Matthew Brown
{24,27}/10/2008

Reading: Koller-Friedman Ch. 10

Space of topics
Outline

**Exact inference as optimization**
Generalized belief propagation
Propagation with approximate messages
Structured variational approximations

General Approach

Represent difficult or intractable distribution \( P_\mathcal{F} \)
with easier distribution \( Q \) chosen from family \( \mathcal{Q} \)

Finding \( Q = \) constrained optimization problem

Note: throughout we assume

\[
P_\mathcal{F}(\mathcal{X}) = \frac{1}{Z} \prod_{\phi \in \mathcal{F}} \phi(\text{scope}_\phi)
\]
KL-Divergence (see also KF 10.1.1)

Challenge: must avoid inference on $P_F$
answer: KL-Divergence

$$\mathbb{D}(P_1 \parallel P_2) = E_{X \sim P_1} \left[ \ln \frac{P_1(X)}{P_2(X)} \right] = \sum_{x \in Val(X)} P_1(x) \ln \frac{P_1(x)}{P_2(x)}$$

Also called relative entropy

$\geq 0$

0 when $P_1 = P_2$

not a distance function - not reflexive: $\mathbb{D}(P_1 \parallel P_2) \neq \mathbb{D}(P_2 \parallel P_1)$

KL-Divergence and exact inference
(see also KF 10.1.1)

Given:
cluster tree $T$ that factorizes $P_F$ and satisfies running
intersection and family preservation,
set of potentials $Q = \{\beta_i : C_i \in T\} \cup \{\mu_{i,j} : (C_i - C_j) \in T\}$
where $C_i =$ clusters in $T$, $S_{i,j} =$ sepsets,
$\beta_i =$ clique potentials, $\mu_{i,j} =$ sepset potentials
with

$$Q(X) = \frac{\prod_{C_i \in T} \beta_i}{\prod_{(C_i - C_j) \in T} \mu_{i,j}}$$

defining a distribution
based on $T$
exact inference (clique tree calibration) is equivalent to
maximizing $-\mathbb{D}(Q \parallel P_F)$ over space of calibrated sets $Q$
KL-Divergence and exact inference
(see also KF 10.1.1)

Constrained optimization:
(See previous slide for term definitions.)

\[
\begin{align*}
\text{CTree-Optimize-KL} & \quad \text{Clique Potentials} \quad \text{Sepset Potentials} \\
\text{Find} & \quad Q = \{\beta_i : C_i \in \mathcal{T}\} \cup \{\mu_{i,j} : (C_i - C_j) \in \mathcal{T}\} \\
\text{that maximize} & \quad -D(Q \| P_{\mathcal{F}}) \\
\text{subject to} & \\
\sum_{C_i} \mu_{i,j}[s_{i,j}] &= \sum_{C_i - C_j} \beta_i[c_i] \\
\forall (C_i - C_j) \in \mathcal{T}, \forall s_{i,j} \in \text{Val}(S_{i,j}) & \\
\sum_{C_i} \beta_i[c_i] &= 1 \quad \forall C_i \in \mathcal{T}.
\end{align*}
\]

Constraints ensure \( Q \):
- is calibrated
- represents valid distribution

\textbf{Theorem:} If \( \mathcal{T} \) is an I-map of \( P_{\mathcal{F}} \), an exact solution exists equivalent to \( P_{\mathcal{F}} \) exists for the constrained optimization problem (on previous slide).
Minimizing KL-Divergence

KL-Divergence awkward to work with
  • sum over all assignments to $\mathcal{X}$.

$$D(P_1 \| P_2) = E_{\mathcal{X} \sim P_1} \left[ \ln \frac{P_1(\mathcal{X})}{P_2(\mathcal{X})} \right] = \sum_{x \in \text{Val}(\mathcal{X})} P_1(x) \ln \frac{P_1(x)}{P_2(x)}$$

Sum over all assignments!

Use energy functional instead (see next slide).

Energy functional (also see KF 10.1.2)

$$D(Q \| P_\mathcal{F}) = \ln Z - F[P_\mathcal{F}, Q]$$

Energy functional:

$$F[P_\mathcal{F}, Q] = \sum_{\phi \in \mathcal{F}} E_{\mathcal{X} \sim Q}[\ln \phi] + H_Q(\mathcal{X})$$

Energy term breaks $P_\mathcal{F}$ into separate factors
  • hopefully factors have small scopes
  • can choose Q to make inference easy
Energy functional proof (also see KF 10.1.2)

\[ D(Q \| P_\mathcal{F}) = \ln Z - F[P_\mathcal{F}, Q] \]
\[ F[P_\mathcal{F}, Q] = \sum_{\phi \in \mathcal{F}} E_{X \sim Q} [\ln \phi] + H_Q(\mathcal{X}) \]
\[ H_Q(\mathcal{X}) = -E_{X \sim Q} [\ln Q(\mathcal{X})] \]

Proof:

\[ D(P_1 \| P_2) = E_{X \sim P_1} \left[ \ln \frac{P_1(\mathcal{X})}{P_2(\mathcal{X})} \right] \]
\[ D(Q \| P_\mathcal{F}) = E_{X \sim Q} [\ln Q(\mathcal{X})] - E_{X \sim Q} [\ln P_\mathcal{F}(X)] \]
\[ \ln P_\mathcal{F}(\mathcal{X}) = \sum_{\phi \in \mathcal{F}} \ln \phi(U_\phi) - \ln Z \]
\[ D(Q \| P_\mathcal{F}) = -H_Q(\mathcal{X}) - E_{X \sim Q} \left[ \sum_{\phi \in \mathcal{F}} \ln \phi(U_\phi) \right] + E_{X \sim Q} [\ln Z] \]
\[ = -F[P_\mathcal{F}, Q] + \ln Z \]

Maximizing energy functional

To find good approximation Q:

- minimize KL-Divergence or equivalently
- maximize energy functional

Use variational methods to find extrema
Factored energy functional (also see KF Def’n 10.2.1)

Given cluster tree $T$, potentials (and messages) $Q$,

$$
\tilde{F}[P_F, Q] = \sum_i F_{C_i \sim \beta_i} \left[ \ln \psi_i \right] + \sum_{C_i \in T} H_{\beta_i}(C_i) - \sum_{(C_i, C_j) \in T} H_{\mu_{i,j}}(S_{i,j})
$$

initial potentials $\psi_i = \prod_{\phi, \alpha(\phi) = i} \phi$,

where $\alpha$ maps $P_F$ factors to $T$ clusters

Useful because all terms are local

- refer to one potential or message

Factored energy functional = energy functional (also see KF 10.2)

**Theorem:**

Given a set $Q$ of calibrated potentials for cluster tree $T$ and

$$
Q(\chi) = \frac{\prod_{C_i \in T} \beta_i}{\prod_{(C_i, C_j) \in T} \mu_{i,j}}
$$

then

$$
\tilde{F}[P_F, Q] = F[P_F, Q]
$$

(For proof see KF Proposition 10.2.2)
Redefinition of optimization problem (KF 10.2)

\begin{equation*}
Q = \{ \beta_i : C_i \in T \} \cup \{ \mu_{i,j} : (C_i - C_j) \in T \}
\end{equation*}

\begin{equation*}
\begin{aligned}
\mu_{i,j}[s_{i,j}] &= \sum_{C_i - C_j} \beta_i [c_i] \\
\sum_{c_i} \beta_i [c_i] &= 1 \\
\beta_i [c_i] &\geq 0
\end{aligned}
\end{equation*}

Constraints ensure \( Q \):
- is calibrated
- represents valid distribution

Finding maxima (also see KF Concept 10.1)

At stationary point (local min, max, or saddle) gradient

\[ \nabla f = \left\langle \frac{\partial f}{\partial \theta_1}, \ldots, \frac{\partial f}{\partial \theta_n} \right\rangle = 0 \]

To find stationary point solve

\[ \frac{\partial}{\partial \theta_1} f(\theta_1, \ldots, \theta_n) = 0 \]

\[ \vdots \]

\[ \frac{\partial}{\partial \theta_n} f(\theta_1, \ldots, \theta_n) = 0 \]

Confirm local maximum (vs. min or saddle)
- Hessian (matrix of 2nd derivatives) is negative definite
Constrained optimization: Lagrange multipliers (also see KF Concept 10.2)

Maximizing $f(\theta)$ subject to constraints $c_1(\theta) = 0$

\[ \cdots \]

$c_m(\theta) = 0$

Equivalent to maximizing Lagrangian (without constraints)

\[
\mathcal{J}(\theta, \lambda) = f(\theta) - \sum_{j=1}^{m} \lambda_j c_j(\theta)
\]

Constrained optimization: Lagrange multipliers (also see KF Concept 10.2)

At Lagrangian’s maximum, constraint curve (red) is tangent to function f’s contours (blue) and orthogonal to f’s gradient

(Images from Wikipedia)
Optimization of factorized energy functional (KF 10.2)

CTree-Optimize

Find that maximize \( Q = \{ \beta_i : C_i \in T \} \cup \{ \mu_{i,j} : (C_i, C_j) \in T \} \)

\[
\tilde{F}[P_F, Q] = \sum_{c_i \in T} \beta_i[c_i]
\]

subject to

\[
\sum_{c_i \in T} \beta_i[c_i] = 1 \quad \forall C_i \in T
\]

\[
\beta_i[c_i] \geq 0 \quad \forall C_i \in T, c_i \in Val(C_i)
\]

Constraints ensure Q:
- is calibrated
- represents valid distribution

Lagrange multipliers for optimization of factorized energy functional (KF 10.2)

Lagrangian

\[
\mathcal{J} = \tilde{F}[P_F, Q] - \sum_{C_i \in T} \lambda_i \left( \sum_{c_i} \beta_i[c_i] - 1 \right)
- \sum_i \sum_{j \in \mathcal{N}_i} \sum_{s_{i,j}} \lambda_{i,j} \left( \sum_{c_i \sim s_{i,j}} \beta_i[c_i] - \mu_{i,j}[s_{i,j}] \right)
\]

\( \mathcal{N}_i \) indicates neighbours in clique tree

\( \text{Next, take derivative of Lagrangian w.r.t. } \beta_i[c_i], \mu_{i,j}[s_{i,j}], \) and \( \lambda \)'s.)
Fixed point equations (also see KF Theorem 10.2.12)

**Theorem:**
Set of potentials $Q$ is a stationary point of CTree-Optimize if and only if

$\exists$ factors $\{\delta_{i\rightarrow j}[S_{i,j}] : C_i \rightarrow C_j \in T\}$ such that

$$
\delta_{i\rightarrow j} \propto \sum_{C_i \rightarrow S_{i,j}} \psi_i \left( \prod_{k \in N_i \setminus \{j\}} \delta_{k \rightarrow i} \right)
$$

fixed point equations

$$
\beta_i \propto \psi_i \left( \prod_{j \in N_i} \delta_{j \rightarrow i} \right)
$$

$$
\mu_{i,j} = \delta_{j \rightarrow i} \cdot \delta_{i \rightarrow j}.
$$

(Proof: see KF Theorem 10.2.12)

Inference as optimization (also see KF 10.2.2)

**Interpretation:**
Optimizing factored energy functional is equivalent to

Doing belief-update message passing (from clique tree lectures) until completion (at a fixed point)

- at fixed point, further application of message passing does not change the potentials any more
Outline

Exact inference as optimization

**Generalized belief propagation**

Propagation with approximate messages
Structured variational approximations

“Loopy” cluster graphs (also see KF Fig 10.1)

Markov net
(Misconception example)

Clique tree

“Loopy” cluster graph

Goal: belief propagation on generalized cluster graphs.
Generalized cluster graph

Definition: Given \( P_\mathcal{F}(\mathcal{X}) = \frac{1}{Z} \prod_{\phi \in \mathcal{F}} \phi(\text{scope}_\phi) \),

a **generalized cluster graph** \( K \) for \( P_\mathcal{F} \) is an undirected graph that satisfies:

- each node associated with a cluster \( C_i \subseteq \mathcal{X} \)
- **family preserving**: \( \forall \phi \in \mathcal{F} \), \( \text{scope}(\phi) = C_i \) for \( \geq 1 \) \( i \)
- each edge \( C_i - C_j \) assoc. w. generalized sepset \( S_{ij} \subseteq C_i \cap C_j \)
  - Note: \( S_{ij} \) can be **subset** of the intersection!

Generalized running intersection property
(also see KF Definition 10.3.1)

Definition: A generalized cluster graph \( K \) satisfies the **generalized running intersection property** if

\( \forall \mathcal{X} \) where \( \mathcal{X} \in C_i \) and \( \mathcal{X} \in C_j \):

- exists path from \( C_i \) to \( C_j \), in which each node contains \( \mathcal{X} \)
- exists only **ONE** such path

Note 1: \( C_i \)'s containing a variable \( \mathcal{X} \) form a tree embedded in \( K \)

Note 2: one path restriction reduces possibility of feedback during belief propagation over loopy graph
  - but if > 1 path between two nodes via different variables, feedback not completely eliminated
Cluster graph examples (also see KF Fig. 10.2)

Generalized calibration (also see KF 10.3.2)

Definition:
A generalized cluster graph with cluster potentials $\beta_i$ is **calibrated** when

$$\sum_{C_i - S_{i,j}} \beta_i = \sum_{C_j - S_{i,j}} \beta_j$$

Note 1: weaker than cluster tree calibration because $S_{i,j} \subseteq C_i \cap C_j$ (as opposed to $S_{ij} = C_i \cap C_j$ for cluster trees)

Note 2: If cluster graph satisfies generalized running intersection property, all clusters containing $X$ agree on marginal of $X$. 
Generalized belief propagation
(also see KF)

How to calibrate generalized cluster graph?
Use belief propagation (a.k.a. message passing)
- similar to clique tree calibration
- either sum-product or belief update message passing
  - (with slight modifications)

Why? Message passing uses local operations
- doesn’t care if graph is tree or loopy
- just have to initialize messages to 1
  - might be no leaf nodes to start things off

---

Generalized sum-product calibration (also see KF Fig. 10.4)

Given set of factors $\mathcal{F}$ and cluster graph $K$,
Initialize $K$:
- assign initial factors $\phi \in \mathcal{F}$ to clusters with $\text{scope}(\phi) \subseteq C_i$
- initial potentials (product of initial factors): $\psi_i \leftarrow \prod_{\phi : \alpha(\phi)=i} \phi$
- messages for each edge $C_i$-$C_j$: $\delta_{i \rightarrow j} \leftarrow 1$
  $\delta_{j \rightarrow i} \leftarrow 1$

While $K$ not calibrated:
pass messages $\delta_{i \rightarrow j}(S_{i,j}) \leftarrow \sum_{C_i - S_{i,j}} \psi_i \cdot \prod_{k \in (N_i-\{j\})} \delta_{k \rightarrow i}$

Final potentials: $\beta_i \leftarrow \psi_i \cdot \prod_{k \in N_i} \delta_{k \rightarrow i}$
Generalized belief update calibration

Belief update calibration can be generalized to cluster graphs similarly to sum-product calibration.

Note: Belief update and sum-product message passing are equivalent for cluster graphs with generalized running intersection property (same as for clique trees).

Generalized belief propagation example (also see KF 10.3.1)

\[ \beta_1[a^0, b^0] \approx \beta_1[a^1, b^1] \gg \beta_1[a^0, b^1] \approx \beta_1[a^1, b^0] \]

and similarity for other \( \beta \)'s

pass messages:

\[ \mu_{1,2}, \mu_{1,2}, \mu_{1,2}, \mu_{1,2}, \mu_{1,2} \]

\( \rightarrow \) damped feedback on \( P(A) \)

• does converge but

• \( P(a^1) \) overestimated
Generalized belief propagation

Two issues:
• not guaranteed to converge (more on this later)
  • if it does converge, may take many passes to do so
• final cluster potentials not necessarily equal to cluster marginals

Advantage
• can be much cheaper than exact inference
  • eg: grid network

Grid network

Compact representation:
node & edge potentials
Exact inference requires
sepsets = cutsets
-> for nxn grid, exact inference
is exponential in n
For loopy belief propagation,
1 round of updating is \(O(n^2)\)
Cluster graph invariance (also see KF 10.3.3.1)

Given generalized cluster graph $K$ over factors $\mathcal{F}$ with cluster potentials $\beta_i$ and sepsets $\mu_{i,j}$ and distribution (un-normalized) $\tilde{P}_\mathcal{F}(\mathcal{X}) = \prod_{\phi \in \mathcal{F}} \phi$

At any iteration of belief update message passing:

$$\tilde{P}_\mathcal{F}(\mathcal{X}) = \frac{\prod_{C_i \in \mathcal{K}} \beta_i[C_i]}{\prod_{(C_i - C_j) \in \mathcal{K}} \mu_{i,j}[S_{i,j}]}$$

Proof: Recall that $\beta_i = \psi_i \prod_{j \in \mathcal{N}_i} \delta_{j \rightarrow i}$ and that $\mu_{i,j} = \delta_{i \rightarrow j} \delta_{j \rightarrow i}$. Thus

$$\frac{\prod_{C_i \in \mathcal{K}} \beta_i[C_i]}{\prod_{(C_i - C_j) \in \mathcal{K}} \mu_{i,j}[S_{i,j}]} = \frac{\prod_{C_i \in \mathcal{K}} \psi_i[C_i] \prod_{j \in \mathcal{N}_i} \delta_{j \rightarrow i} \delta_{i \rightarrow j}}{\prod_{(C_i - C_j) \in \mathcal{K}} \delta_{i \rightarrow j} \delta_{j \rightarrow i}}$$

$$= \prod_{C_i \in \mathcal{K}} \psi_i[C_i]$$

$$= \prod_{\phi \in \mathcal{F}} \phi(u_\phi) = \tilde{P}_\mathcal{F}(\mathcal{X})$$

$\delta_{i \rightarrow j}$ appears exactly once in numerator and denominator.
Cluster graphs and marginals (also see KF 10.3.3.2)

Even though

\[ \tilde{P}_F(\mathcal{X}) = \frac{\prod_{C_i \in \mathcal{K}} \beta_i[C_i]}{\prod_{(C_i - C_j) \in \mathcal{K}} \mu_{i,j}[S_{i,j}]} \]

we can show \( \tilde{P}_F(C_i) \neq \beta_i(C_i) \)

(Example on next slide)
Generalized belief propagation and convergence
(also see KF 10.3.4)

Convergence not guaranteed, generally.
Oscillation or divergence possible
\( \alpha \)-contraction does guarantee convergence
non-trivial to demonstrate \( \alpha \)-contraction

heuristics:
max # iterations
threshold on change of cluster graph measure

Building cluster graphs (also see KF example 10.3.6)

Propagates only C’s influence
Propagates influence of B & C & BC
Building cluster graphs (also see KF example 10.3.6)

Including more representation for variable interactions in cluster graph structure:

- more accurate
  - increases what information gets passed around
- harder to ensure running intersection property

Cluster graph for grid network
(also see KF Fig. 10.7)

Cluster graph for 3x3 grid network

Special name for grid inference: loopy belief propagation
Bethe approximation (also see KF Fig. 10.8)

Large clusters: scope(\(\phi\))

Small clusters: individual X’s

Satisfies:
- family preservation
- running intersection property

Disadvantage: interactions between X’s lost during propagation through small clusters

---

Interactions among variables (also see KF 10.3.5.3)

Extra node to address B,C interactions

Can we apply this idea to all variable pairs? (see next slide)
Interactions among variables (also see KF 10.3.5.3)

All pairs represented but violates running intersection property (look at C).

C removed from Node 5 to satisfy running intersection property.

Recall optimization of factored energy functional

Recall: exact inference equivalent to optimizing factored energy functional subject to some constraints.

CTree-Optimize

Find that maximize

\[ Q = \{ \beta_i : C_i \in T \} \cup \{ \mu_{i,j} : (C_i - C_j) \in T \} \]

\[ \tilde{F}[P_{\mathcal{T}}, Q] \]

subject to

\[ \mu_{i,j}[s_{i,j}] = \sum_{c_i - s_{i,j}} \beta_i[c_i] \quad \forall (C_i - C_j) \in T, \forall s_{i,j} \in Val(S_{i,j}) \]

\[ \sum_{C_i} \beta_i[c_i] = 1 \quad \forall C_i \in T \]

\[ \beta_i[c_i] \geq 0 \quad \forall C_i \in T, c_i \in Val(C_i) \]
Variational analysis of generalized BP (also see KF 10.3.6)

Constrained optimization view of generalized belief propagation:

approximation #1: \[ \tilde{\mathcal{F}}[P_F, Q] \approx \mathcal{F}[P_F, Q] \]

- factored energy functional
- energy functional

approximation #2: optimize over pseudo-marginals
- superspace of PDFs that factorize of cluster graph \( K \)
- recall: \( \beta_i(C_i) \approx \tilde{P}_F(C_i) \)
- allows one to model complex distribution with simpler graph - eg: smaller clusters
  - computing on cluster potentials is exponential in size of cluster scope

Pseudo-marginals (also see KF 10.3.6)

approximation #2: optimize over pseudo-marginals
(i.e. superspace of PDFs that factorize of cluster graph \( K \))

want to optimize \( Q_P = \{ P(C_i) \}_{C_i \in K} \cup \{ P(S_{i,j}) \}_{(C_i, C_j) \in K} \)
- but exist calibrated \( Q_P \) that do not represent coherent probability function
- NP-hard to determine coherency for given \( Q_P \)

so, optimize over locally consistent subset in which

\[
\begin{align*}
\mu_{i,j}[s_{i,j}] &= \sum_{C_i - C_j} \beta_i[c_i] \quad \forall (C_i - C_j) \in K, \forall s_{i,j} \in \text{Val}(S_{i,j}) \\
\sum_{C_i} \beta_i[c_i] &= 1 \quad \forall C_i \in K \\
\beta_i[c_i] &\geq 0 \quad \forall C_i \in K, c_i \in \text{Val}(C_i)(C)
\end{align*}
\]
Fixed point equations for generalized BP
(also see KF Theorem 10.3.7 & corollaries)

**Theorem:**
Set of potentials $Q$ is a stationary point of CGraph-Optimize if and only if

$\exists$ factors $\{\delta_{i\rightarrow j}[S_{i,j}] : C_i \rightarrow C_j \in T\}$ such that

\[
\delta_{i\rightarrow j} \propto \sum_{C_i \rightarrow S_{i,j}} \psi_i \left( \prod_{k \in \mathcal{N}_i - \{j\}} \delta_{k\rightarrow i} \right)
\]

fixed point equations

\[
\beta_i \propto \psi_i \left( \prod_{j \in \mathcal{N}_i} \delta_{j\rightarrow i} \right)
\]

$\mu_{i,j} = \delta_{j\rightarrow i} \cdot \delta_{i\rightarrow j}.$

(Proof: see KF Theorem 10.3.7)

---

Stationary points of generalized belief propagation
(also see KF Theorem 10.3.7 & corollaries)

**Corollary:**
Finding convergence point of cluster graph belief propagation (either sum-product of belief update)

is equivalent to

finding stationary point of factored energy functional
Convergence

- maxima not necessarily fixed points
  - boundary points
- maxima not the only fixed points
  - minima, saddles
- unstable maxima
  - even if BP could reach maximum, slight perturbation causes divergence or oscillation

Outline

Exact inference as optimization
Generalized belief propagation

**Propagation with approximate messages**
Structured variational approximations
Propagation with approximate messages

Instead of approximating distribution (as above), use complicated graph structure
- larger cluster scopes
- approximate messages
  - maintains tractability

For full discussion, see KF section 10.4

Outline

Exact inference as optimization
Generalized belief propagation
Propagation with approximate messages

Structured variational approximations
Structure variational approximations

Find $Q \in \mathcal{Q}$ that optimizes exact energy functional $F[P_{\mathcal{F}}, Q]$

Choose family $\mathcal{Q}$ to make computation tractable
- tradeoff between tractability vs. accuracy in selection of $\mathcal{Q}$

Note: approach guaranteed to converge

Mean field approximation (also see KF 10.5.1.1)

Choose $\mathcal{Q}$ such that $Q(\mathbf{x}) = \prod_{i} Q(X_i)$
- product of independent marginals
- computationally attractive
- very restrictive - loses much information in $P_{\mathcal{F}}$

Approach:
- derive fixed point equations
- build iterative algorithm from them
Mean field makes energy functional nice  
(also see KF 10.5.1.1)

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

\[ Q(\mathcal{X}) = \prod_{i} Q(X_i) \] implies:

1)  
\[ E_u \sim Q[\ln \phi] = \sum_{u_\phi} Q(u_\phi) \ln \phi(u_\phi) \]

\[ = \sum_{u_\phi} \left( \prod_{X_i \in U_\phi} Q(x_i) \right) \ln \phi(u_\phi) \]

2)  
\[ H_Q(\mathcal{X}) = \sum_{i} H_Q(X_i) \]

So, Energy functional = sum of local terms

Complexity depends of factor sizes

• not network topology

Maximizing energy functional (see KF 10.5.1.2)

Method of Lagrange multipliers

\[ Q(x_i) = \frac{1}{Z_i} \exp \left\{ \sum_{\phi \in \mathcal{F}} E_{\mathcal{X} \sim Q}[\ln \phi \mid x_i] \right\} \]

Can show:  
\[ Q(x_i) = \frac{1}{Z_i} \exp \{ E_{X_{-i} \sim Q}[\ln P_{\mathcal{F}}(x_i \mid X_{-i})] \} \]

i.e. marginal on \( x_i \) = geometric mean of conditional probability of \( x_i \) given all other variables

\[ GeometricMean(a_1, a_2, ..., a_n) = \sqrt[n]{\prod_{i=1}^{n} a_i} \]
Maximizing energy functional (see KF 10.5.1.3 & Fig. 10.21)

Method of Lagrange multipliers

-> fixed point equations

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

Can show:

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

i.e. only sum across factors whose scope includes \( X_i \)

-> can compute optimal \( Q(X_i) \) given other marginals

-> Mean-Field algorithm: iterative coordinate ascent optimizing one \( Q(X_i) \) at a time

Mean field algorithm (also see KF Fig. 10.21)

**Procedure** Mean-Field (\( \mathcal{F}, Q_0 \))

\[ Q \leftarrow Q_0 \]

\( Unprocessed \leftarrow X \)

while \( Unprocessed \neq \emptyset \)

Choose \( X_i \) from \( Unprocessed \)

\[ Q_{old}(X_i) \leftarrow Q(X_i) \]

for \( x_i \in \text{Val}(X_i) \) do

\[ Q(x_i) \leftarrow \exp \left\{ \sum_{\phi: X_i \in \text{Scope}[\phi]} E_{(U_\phi \sim \{X_i\}) \sim Q[\ln \phi(U_\phi, x_i)]} \right\} \]

Normalize \( Q(X_i) \) to sum to one

if \( Q_{old}(X_i) \neq Q(X_i) \) then

\( Unprocessed \leftarrow Unprocessed \cup (\cup_{\phi: X_i \in \text{Scope}[\phi]} \text{Scope}[\phi]) \)

\( Unprocessed \leftarrow Unprocessed - \{X_i\} \)

return \( Q \)
Maximizing energy functional (see KF Theorem 10.5.8)

Mean-Field algorithm:
- guaranteed to converge
- optimal distribution $Q^*$ returned is also stationary point of $E[P, Q]$ as long as $Q(x) = \prod_i Q(x_i)$ is a distribution
- $Q^*$ could in theory be saddle point or local minimum (rather than local maximum)
  - BUT saddles and minima are unstable optima
  - in practice, algorithm converges to local maximum
  - Note: not necessarily global maximum

Mean field approximation

Summary:
- nice computational properties
- limited ability to represent distributions
  - all variables assumed independent of each other
  - eg: cannot represent XOR
Markov network approximations (also see KF 10.5.2)

Let $\mathcal{Q}$ be subset of Gibbs distributions:

$$Q(\mathcal{X}) = \frac{1}{Z_\mathcal{Q}} \prod_{j=1}^{J} \psi_j \quad \text{Scope}[\psi_j] = C_j$$

much more expressive than mean field distribution

Markov network approximations (also see KF 10.5.2)

Must choose structure for Markov net for $\mathcal{Q}$
examples for grid network

Preserve only some dependencies
Large computational simplification
Energy functional

\[ Q(\mathcal{X}) = \frac{1}{Z_Q} \prod_{j=1}^{J} \psi_j \]

gives us:

\[ E_{U_{\phi} \sim Q}[\ln \phi] \]

computed efficiently using exact inference

\[ H_Q(\mathcal{X}) = -\sum_{j=1}^{J} E_{U_{\psi_j} \sim Q}[\ln \psi_j(C_j)] + \ln Z_Q \]

Energy functional

\[ F[P_\mathcal{F}, Q] = \sum_{k=1}^{K} E_{U_{\phi_k} \sim Q}[\ln \phi_k] - \sum_{j=1}^{J} E_{U_{\psi_j} \sim Q}[\ln \psi_j(C_j)] + \ln Z_Q \]

Fixed point equations (also see KF Theorem 10.5.12 & corollary)

At stationary point of energy functional

\[ \psi_j(c_j) \propto \exp \left\{ E_Q[\ln \tilde{P}_\mathcal{F} \mid c_j] - \sum_{k \neq j} E_Q[\ln \psi_k \mid c_j] \right\} \]

Leads to iterative individual factor-wise maximization

- \( \psi_j \) optimized assuming values for all other \( \psi_k \)
- computing expectations non-trivial (unlike mean field)
  - requires inference over Markov net for multiple assignments of factor \( \psi_j \)
Sequential vs. parallel updating

**Sequential:**

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

Update on factor $\psi_j$ at a time

Difficulty: inference required for each step
- network $H_Q$ reparameterized for each assignment $\psi_j$
- if using clique tree inference, recalibrate in each step

**Parallel:**

Compute right hand side of fixed point equation for each $\psi_j$ simultaneously
- requires only single calibration step for each iteration

Then re-estimate all potentials $\psi_j$
Repeat

Sequential vs. parallel updating

**Sequential:**

Slower
Guaranteed to converge to a local maximum

**Parallel:**

Often much faster, especially with complicated $H_Q$
No convergence guarantees
The computed optimal $\psi_j$ given all other $\psi_k$ may not be optimal for new assignments of other $\psi_k$
Further material (see KF 10.5.2.3, 10.5.2.4)

Don’t have time to cover in class:
  Simplifying update equations
  Simplifying family $Q$