Prof. Jeff Bilmes

University of Washington, Seattle
Department of Electrical Engineering
http://melodi.ee.washington.edu/~bilmes

Nov 24th, 2014
Announcements


- Should have read chapters 1, 2, 3, 4 in this book. Read chapter 5.

- Assignment due Wednesday (Nov 26th) night, 11:45pm. Final project proposal updates and progress report (one page max).
Class Road Map - EE512a

- L1 (9/29): Introduction, Families, Semantics
- L2 (10/1): MRFs, elimination, Inference on Trees
- L3 (10/6): Tree inference, message passing, more general queries, non-tree
- L4 (10/8): Non-trees, perfect elimination, triangulated graphs
- L5 (10/13): triangulated graphs, $k$-trees, the triangulation process/heuristics
- L6 (10/15): multiple queries, decomposable models, junction trees
- L7 (10/20): junction trees, begin intersection graphs
- L8 (10/22): intersection graphs, inference on junction trees
- L9 (10/27): inference on junction trees, semirings,
- L10 (11/3): conditioning, hardness, LBP
- L11 (11/5): LBP, exponential models,
- L12 (11/10): exponential models, mean params and polytopes,
- L13 (11/12): polytopes, tree outer bound, Bethe entropy approx.
- L14 (11/17): Bethe entropy approx, loop series correction
- L15 (11/19): Hypergraphs, posets, Mobius, Kikuchi
- L16 (11/24):
- L17 (11/26):
- L18 (12/1):
- L19 (12/3):
- Final Presentations: (12/10):

Finals Week: Dec 8th-12th, 2014.
Drawing/Visualizing Hypergraphs as Bipartite Graphs

- Hypergraph (shaded regions) on left, while bipartite graph representation on the right.
Hypergraph, edge representations

- It is possible to represent hypergraphs by only showing their hyperedges.

- Here, we see graphical representations of three hypergraphs. Subsets of nodes corresponding to hyperedges are shown in rectangles, whereas the arrows represent inclusion relations among hyperedges.

- Which ones, if any, are in reduced representation?
Möbius Inversion Lemma and Inclusion-Exclusion

- For any \( A \subseteq V \), define two functions \( \Omega : 2^V \to \mathbb{R} \) and \( \Upsilon : 2^V \to \mathbb{R} \).
- Then the above inclusion-exclusion principle is one instance of the more general Möbius Inversion lemma, namely that each of the below two equations implies the other.

\[
\forall A \subseteq V : \Upsilon(A) = \sum_{B : B \subseteq A} \Omega(B) \quad (16.13)
\]

\[
\forall A \subseteq V : \Omega(A) = \sum_{B : B \subseteq A} (-1)^{|A \setminus B|} \Upsilon(B) \quad (16.14)
\]

- Möbius Inversion lemma is also used to prove the Hammersley-Clifford theorem (that factorization and Markov property definitions of families are identical for positive distributions).
- We use it here to come up with alternative expressions for the entropy and for the marginal polytope.
Möbius Inversion Lemma for posets

- Let $\mathcal{P}$ be a partially ordered set with binary relation $\leq$.
- A zeta function of a poset is a mapping $\zeta : \mathcal{P} \times \mathcal{P} \to \mathbb{R}$ defined by
\[
\zeta(g, h) = \begin{cases} 
1 & \text{if } g \leq h, \\
0 & \text{otherwise.}
\end{cases}
\tag{16.23}
\]
- The Möbius function $\omega : \mathcal{P} \times \mathcal{P} \to \mathbb{R}$ is the multiplicative inverse of this function. It is defined recursively:
  - $\omega(g, g) = 1$ for all $g \in \mathcal{P}$
  - $\omega(g, h) = 0$ for all $h : h \nleq g$.
  - Given $\omega(g, f)$ defined for $f$ such that $g \leq f < h$, we define
\[
\omega(g, h) = -\sum_{\{f|g \leq f < h\}} \omega(g, f)
\tag{16.24}
\]
- Then, $\omega$ and $\zeta$ are multiplicative inverses, in that
\[
\sum_{f \in \mathcal{P}} \omega(g, f)\zeta(f, h) = \sum_{\{f|g \leq f \leq h\}} \omega(g, f) = \delta(g, h)
\tag{16.25}
\]
Lemma 16.2.8 (General Möbius Inversion Lemma)

Given real valued functions \( \Upsilon \) and \( \Omega \) defined on poset \( P \), then \( \Omega(h) \) may be expressed via \( \Upsilon(\cdot) \) via

\[
\Omega(h) = \sum_{g \preceq h} \Upsilon(g) \quad \text{for all } h \in P \tag{16.23}
\]

iff \( \Upsilon(h) \) may be expressed via \( \Omega(\cdot) \) via

\[
\Upsilon(h) = \sum_{g \preceq h} \Omega(g) \omega(g, h) \quad \text{for all } h \in P \tag{16.24}
\]

When \( P = 2^V \) for some set \( V \) (so this means that the poset consists of sets and all subsets of an underlying set \( V \)) this can be simplified, where \( \preceq \) becomes \( \subseteq \); and \( \succeq \) becomes \( \supseteq \), like we saw above. (see Stanley, “Enumerative Combinatorics” for more info.)
Back to Kikuchi: Möbius and expressions of factorization

- Suppose we are given marginals that factor w.r.t. a hypergraph $G = (V, E)$, so we have $\mu = (\mu_h, h \in E)$, then we can define new functions $\varphi = (\varphi_h, h \in E)$ via Möbius inversion lemma as follows

$$ \log \varphi_h(x_h) \triangleq \sum_{g \preceq h} \omega(g, h) \log \mu_g(x_g) \quad (16.23) $$

- From Möbius inversion lemma, this then gives us a new way to write the log marginals, i.e., as

$$ \log \mu_h(x_h) = \sum_{g \preceq h} \log \varphi_g(x_g) \quad (16.24) $$

- Key, when $\varphi_h$ is defined as above, and $G$ is a hypertree we have

$$ p_\mu(x) = \prod_{h \in E} \varphi_h(x_h) \quad (16.25) $$

$\Rightarrow$ general way to factorize a distribution that factors w.r.t. a hypergraph.
multi-information decomposition

- Using Möbius, and Eqn. (??) we can write

\[ I_h(\mu_h) = \sum_{x_h} \mu_h(x_h) \log \varphi_h(x_h) = \sum_{x_h} \mu_h(x_h) \left( \sum_{g \preceq h} \omega(g, h) \log \mu_g(x_g) \right) \]

\[ = \sum_{g \preceq h} \omega(g, h) \left\{ \sum_{x_h} \mu_h(x_h) \log \mu_g(x_g) \right\} \]

\[ = \sum_{f \preceq h} \sum_{e \succeq f} \omega(f, e) \left\{ \sum_{x_f} \mu_f(x_f) \log \mu_f(x_f) \right\} = -\sum_{f \preceq h} c(f) H_f(\mu_f) \]

where we define overcounting numbers (\(\sim\) shattering coefficient)

\[ c(f) \triangleq \sum_{e \succeq f} \omega(f, e) \quad (16.31) \]

- This gives us a new expression for the hypertree entropy

\[ H_{\text{hyper}}(\mu) = \sum_{h \in E} c(h) H_h(\mu_h) \quad (16.32) \]
Usable to get Kikuchi variational approximation

- Sum to one constraint:

\[
\sum_{x_h} \tau_h(x_h) = 1 \tag{16.33}
\]

- Local agreement via the hypergraph constraint. For any \( g \preceq h \) must have marginalization condition

\[
\sum_{x_h \setminus g} \tau_h(x_h) = \tau_g(x_g) \tag{16.34}
\]

- Define new polyhedral constraint set \( \mathbb{I}_t(G) \)

\[
\mathbb{I}_t(G) = \{ \tau \geq 0 \mid \text{Equations (16.47) } \forall h, \text{ and (16.55) } \forall g \preceq h \text{ hold} \} \tag{16.35}
\]
Kikuchi variational approximation, entropy approx

- Generalized approximate (app) entropy for the hypergraph:

\[ H_{\text{app}} = \sum_{g \in E} c(g)H_g(\tau_g) \]  \hspace{1cm} (16.33)

where \( H_g \) is hyperedge entropy and overcounting number defined by:

\[ c(g) = \sum_{f \supseteq g} \omega(g, f) \]  \hspace{1cm} (16.34)
Variational Approach Amenable to Approximation

- Original variational representation of log partition function

\[
A(\theta) = \sup_{\mu \in \mathcal{M}} \{ \langle \theta, \mu \rangle - A^*(\mu) \}
\]  \hspace{1cm} (16.1)

where dual takes form:

\[
A^*(\mu) = \sup_{\theta \in \Omega} (\langle \theta, \mu \rangle - A(\theta)) = \begin{cases} 
- H(p_{\theta(\mu)}) & \text{if } \mu \in \mathcal{M}^\circ \\
+\infty & \text{if } \mu \notin \overline{\mathcal{M}} 
\end{cases}
\]  \hspace{1cm} (16.2)
Variational Approach Amenable to Approximation

- Original variational representation of log partition function

\[
A(\theta) = \sup_{\mu \in \mathcal{M}} \{ \langle \theta, \mu \rangle - A^*(\mu) \} \tag{16.1}
\]

where dual takes form:

\[
A^*(\mu) = \sup_{\theta \in \Omega} (\langle \theta, \mu \rangle - A(\theta)) = \begin{cases} 
-\mathcal{H}(p_{\theta(\mu)}) & \text{if } \mu \in \mathcal{M}^\circ \\
+\infty & \text{if } \mu \notin \overline{\mathcal{M}}
\end{cases} \tag{16.2}
\]

- Given efficient expression for \( A(\theta) \), we can compute marginals of interest.
Variational Approach Amenable to Approximation

- Original variational representation of log partition function

\[ A(\theta) = \sup_{\mu \in \mathcal{M}} \{ \langle \theta, \mu \rangle - A^*(\mu) \} \]  

(16.1)

where dual takes form:

\[ A^*(\mu) = \sup_{\theta \in \Omega} (\langle \theta, \mu \rangle - A(\theta)) = \begin{cases} -H(p_{\theta(\mu)}) & \text{if } \mu \in \mathcal{M}^\circ \\ +\infty & \text{if } \mu \notin \overline{\mathcal{M}} \end{cases} \]  

(16.2)

- Given efficient expression for \( A(\theta) \), we can compute marginals of interest.

- Above expression (dual of the dual) offers strategies to approximate or (upper or lower) bound \( A(\theta) \). We either approximate \( \mu \) or \(-A^*(\mu)\) or (most likely) both.
Variational Approximations we cover

1. Set $\mathcal{M} \leftarrow \mathbb{L}$ and $-A^*(\mu) \leftarrow H_{\text{Bethe}}(\tau)$ to get Bethe variational approximation, LBP fixed point.
Variational Approximations we cover

1. Set \( \mathcal{M} \leftarrow \mathbb{L} \) and \( -A^*(\mu) \leftarrow H_{\text{Bethe}}(\tau) \) to get **Bethe variational approximation**, LBP fixed point.

2. Set \( \mathcal{M} \leftarrow \mathbb{L}_t(G) \) (hypergraph marginal polytope), \( -A^*(\mu) \leftarrow H_{\text{app}}(\tau) \) where \( H_{\text{app}} = \sum_{g \in E} c(g)H_g(\tau_g) \) (via Möbius) to get **Kikuchi variational approximation**, message passing on hypergraphs.
Kikuchi variational approximation

This at last gets the Kikuchi variational approximation

\[ A_{\text{Kikuchi}}(\theta) = \max_{\tau \in \mathcal{L}_t(G)} \{ \langle \theta, \tau \rangle + H_{\text{app}}(\tau) \} \] (16.1)
Kikuchi variational approximation

- This at last gets the Kikuchi variational approximation

\[
A_{\text{Kikuchi}}(\theta) = \max_{\tau \in \mathbb{L}_t(G)} \{ \langle \theta, \tau \rangle + H_{\text{app}}(\tau) \}
\]  

(16.1)

- For a graph, this is exactly \( A_{\text{Bethe}}(\theta) \).
Kikuchi variational approximation

- This at last gets the Kikuchi variational approximation

\[ A_{Kikuchi}(\theta) = \max_{\tau \in \mathbb{L}_t(G)} \{ \langle \theta, \tau \rangle + H_{app}(\tau) \} \]  (16.1)

- For a graph, this is exactly \( A_{Bethe}(\theta) \).

- Also, if hypergraph is junction tree (r.i.p. holds, tree-local consistency implies global consistency), then also exact (although expensive, exponential in the tree-width to compute \( H_{app} \)).
Kikuchi variational approximation

- This at last gets the Kikuchi variational approximation

\[
A_{\text{Kikuchi}}(\theta) = \max_{\tau \in \mathbb{L}_t(G)} \{ \langle \theta, \tau \rangle + H_{\text{app}}(\tau) \} \tag{16.1}
\]

- For a graph, this is exactly \( A_{\text{Bethe}}(\theta) \).
- Also, if hypergraph is junction tree (r.i.p. holds, tree-local consistency implies global consistency), then also exact (although expensive, exponential in the tree-width to compute \( H_{\text{app}} \)).
- We can define message passing algorithms on the hypertree, and show that if it converges, it is a fixed point of the associated Lagrangian.
Kikuchi variational approximation, 3x3 grid example

- Example, left is 3x3 grid, right is optimal junction tree cover.

![Diagram showing a 3x3 grid and its optimal junction tree cover.](image-url)
Kikuchi variational approximation, 3x3 grid example

- Example, left is 3x3 grid, right is optimal junction tree cover.

- Treewidth is 4, so complexity is $O(r^5)$. 
Kikuchi variational approximation, 3x3 grid example

- Example, left is 3x3 grid, right is optimal junction tree cover.

- Treewidth is 4, so complexity is $O(r^5)$.

- In general, for $n \times n$ grid structured graph, treewidth is $O(n)$ (grows as the square root of the number of nodes).
Kikuchi variational approximation, 3x3 grid example

- Left is clustering of vertices in 3x3 grid, and right is hyperedge graph/region graph.

Complexity is only $O(r^4)$ and will stay $O(r^4)$ even as $n$ gets bigger.
Kikuchi variational approximation, 3x3 grid example

- Left is clustering of vertices in 3x3 grid, and right is hyperedge graph/region graph.

- Complexity is only $O(r^4)$ and will stay $O(r^4)$ even as $n$ gets bigger (since clusters are at most size four).
Generalized BP (GBP): Key idea

- Key idea: sets of nodes send messages to other sets of nodes.
Generalized BP (GBP): Key idea

- Key idea: sets of nodes send messages to other sets of nodes.
- The node sets that communicate with each other represented using hypergraph (hyperedges are the node sets)

\[ \text{Kikuchi}(\theta) = \max_{\tau \in \mathcal{L}(G)} \left\{ \langle \theta, \tau \rangle + H_{\text{app}}(\tau) \right\} \] (16.2)
Generalized BP (GBP): Key idea

- Key idea: sets of nodes send messages to other sets of nodes.
- The node sets that communicate with each other represented using hypergraph (hyperedges are the node sets)
- Standard LBP algorithm is merely a special case of GBP
Generalized BP (GBP): Key idea

- Key idea: sets of nodes send messages to other sets of nodes.
- The node sets that communicate with each other represented using hypergraph (hyperedges are the node sets)
- Standard LBP algorithm is merely a special case of GBP
- Different choices of node sets/hyperedges and message passings give different GBP algorithms.
Generalized BP (GBP): Key idea

- Key idea: sets of nodes send messages to other sets of nodes.
- The node sets that communicate with each other represented using hypergraph (hyperedges are the node sets)
- Standard LBP algorithm is merely a special case of GBP
- Different choices of node sets/hyperedges and message passing give different GBP algorithms.
- This gives the user a gradual tradeoff between the most expensive, intractable, and accurate junction tree algorithm, and the least expensive but possibly quite inaccurate LBP algorithm.
Key idea: sets of nodes send messages to other sets of nodes.

The node sets that communicate with each other represented using hypergraph (hyperedges are the node sets).

Standard LBP algorithm is merely a special case of GBP.

Different choices of node sets/hyperedges and message passings give different GBP algorithms.

This gives the user a gradual tradeoff between the most expensive, intractable, and accurate junction tree algorithm, and the least expensive but possibly quite inaccurate LBP algorithm.

Allows a trade-off between complexity for accuracy!
Generalized BP (GBP): Key idea

- Key idea: sets of nodes send messages to other sets of nodes.
- The node sets that communicate with each other represented using hypergraph (hyperedges are the node sets).
- Standard LBP algorithm is merely a special case of GBP.
- Different choices of node sets/hyperedges and message passings give different GBP algorithms.
- This gives the user a gradual tradeoff between the most expensive, intractable, and accurate junction tree algorithm, and the least expensive but possibly quite inaccurate LBP algorithm.
- Allows a trade-off between complexity for accuracy!
- In many cases, convergence of GBP will be at fixed points of the Lagrangian for the generalized variational approximation.

\[
A_{\text{Kikuchi}}(\theta) = \max_{\tau \in \mathcal{L}_t(G)} \{ \langle \theta, \tau \rangle + H_{\text{app}}(\tau) \}
\]
GBP examples: parent-to-child

In hypergraph Hasse-like diagram, arrows point from parent (superset) to child (subset). Ex: on the right, set \{1, 2, 4, 5\} is the parent of both \{2, 5\} and \{4, 5\}. 
GBP examples: parent-to-child

In hypergraph Hasse-like diagram, arrows point from parent (superset) to child (subset). Ex: on the right, set \{1, 2, 4, 5\} is the parent of both \{2, 5\} and \{4, 5\}.

- For \( h \in E \), let \( \text{Par}(h) \) be the set of parents. Also define descendants as \( \mathcal{D}(h) = \{ g \in E | g \prec h \} \) and ancestors as \( \mathcal{A}(h) = \{ g \in E | g \succ h \} \).
GBP examples: parent-to-child

In hypergraph Hasse-like diagram, arrows point from parent (superset) to child (subset). Ex: on the right, set \{1, 2, 4, 5\} is the parent of both \{2, 5\} and \{4, 5\}.

- For \( h \in E \), let \( \text{Par}(h) \) be the set of parents. Also define descendants as \( \mathcal{D}(h) = \{ g \in E | g \prec h \} \) and ancestors as \( \mathcal{A}(h) = \{ g \in E | g \succ h \} \).

- Also define \( \mathcal{D}^+(h) = \mathcal{D}(h) \cup \{ h \} \) and \( \mathcal{A}^+(h) = \mathcal{A}(h) \cup \{ h \} \).
GBP examples: parent-to-child

In hypergraph Hasse-like diagram, arrows point from parent (superset) to child (subset). Ex: on the right, set \{1, 2, 4, 5\} is the parent of both \{2, 5\} and \{4, 5\}.

- For \( h \in E \), let \( \text{Par}(h) \) be the set of parents. Also define descendants as \( \mathcal{D}(h) = \{ g \in E | g \prec h \} \) and ancestors as \( \mathcal{A}(h) = \{ g \in E | g \succ h \} \).
- Also define \( \mathcal{D}^+(h) = \mathcal{D}(h) \cup \{ h \} \) and \( \mathcal{A}^+(h) = \mathcal{A}(h) \cup \{ h \} \).
- If \( f \succ g \) then \( x_f \) has more variables than \( x_g \) and one can perform a message of the form \( M_{f \rightarrow g}(x_g) = \sum_{f \setminus g} \tau(x_f) = \sum_{f \setminus g} \tau(x_g, x_{f \setminus g}) \).
GBP examples: parent-to-child message

Then parent-to-child message passing takes the form:

\[
\tau_h(x_h) \propto \prod_{g \in \mathcal{D}^+(h)} \exp(\theta(x_g)) \prod_{g \in \mathcal{D}^+(h)} \prod_{f \in \text{Par}(g) \setminus \mathcal{D}^+(h)} M_{f \to g}(x_g)
\]

(16.3)
GBP examples: parent-to-child message

Then parent-to-child message passing takes the form:

$$\tau_h(x_h) \propto \left[ \prod_{g \in \mathcal{D}^+(h)} \exp(\theta(x_g)) \right] \left[ \prod_{g \in \mathcal{D}^+(h)} \prod_{f \in \text{Par}(g) \setminus \mathcal{D}^+(h)} M_{f \rightarrow g}(x_g) \right]$$  \hspace{1cm} (16.3)

We form marginal at $h$

- from the factors associated with each hyperedge, namely $\exp(\theta(x_g))$, and by the messages sent to $h$ and $h$’s descendants from other parents.
Consider message for hyperedge $h = \{1, 2, 4, 5\}$, which has factors $\psi'$ associated with (regular graph) edges $\{1, 2\}, \{2, 5\}, \{4, 5\}$, and $\{1, 4\}$ and also unary factors for each of the nodes 1, 2, 4, and 5 (eg., to associate evidence into the model).
 GBP examples: parent-to-child message, grid graph

Consider message for hyperedge $h = \{1, 2, 4, 5\}$, which has factors $\psi'$ associated with (regular graph) edges $\{1, 2\}$, $\{2, 5\}$, $\{4, 5\}$, and $\{1, 4\}$ and also unary factors for each of the nodes 1, 2, 4, and 5 (eg., to associate evidence into the model).

Then $D^+(h) = \{\{1, 2, 4, 5\}, \{4, 5\}, \{2, 5\}, \{5\}\}$. 
Consider message for hyperedge \( h = \{1, 2, 4, 5\} \), which has factors \( \psi' \) associated with (regular graph) edges \( \{1, 2\} \), \( \{2, 5\} \), \( \{4, 5\} \), and \( \{1, 4\} \) and also unary factors for each of the nodes 1, 2, 4, and 5 (eg., to associate evidence into the model).

Then \( D^+(h) = \{\{1, 2, 4, 5\}, \{4, 5\}, \{2, 5\}, \{5\}\} \).

We get an expression for the marginal at \( h \) using the above formula.

\[
\tau_{1,2,4,5} \propto \psi'_{1,2} \psi'_{1,4} \psi'_{2,5} \psi'_{4,5} \psi'_{1} \psi'_{2} \psi'_{4} \psi'_{5} \\
\times M_{\{2,3,5,6\} \rightarrow \{2,5\}} M_{\{4,5,7,8\} \rightarrow \{4,5\}} M_{\{5,6\} \rightarrow \{5\}} M_{\{5,8\} \rightarrow \{5\}}
\] (16.4)
Consider message for hyperedge \( h = \{1, 2, 4, 5\} \), which has factors \( \psi' \) associated with (regular graph) edges \( \{1, 2\} \), \( \{2, 5\} \), \( \{4, 5\} \), and \( \{1, 4\} \) and also unary factors for each of the nodes 1, 2, 4, and 5 (eg., to associate evidence into the model).

Then \( D^+(h) = \{\{1, 2, 4, 5\}, \{4, 5\}, \{2, 5\}, \{5\}\} \).

We get an expression for the marginal at \( h \) using the above formula.

\[
\tau_{1,2,4,5} \propto \psi'_{1,2} \psi'_{1,4} \psi'_{2,5} \psi'_{4,5} \psi'_{1} \psi'_{2} \psi'_{4} \psi'_{5} \\
\times M_{\{2,3,5,6\} \to \{2,5\}} M_{\{4,5,7,8\} \to \{4,5\}} M_{\{5,6\} \to \{5\}} M_{\{5,8\} \to \{5\}} \tag{16.4}
\]

This could repeat for each of the largest clusters, until convergence.
Theorem 16.4.3 (Relationship between $A$ and $A^*$)

(a) For any $\mu \in \mathcal{M}^\circ$, $\theta(\mu)$ unique canonical parameter sat. matching condition, then conj. dual takes form:

$$
A^*(\mu) = \sup_{\theta \in \Omega} (\langle \theta, \mu \rangle - A(\theta)) = \begin{cases} 
-\mathcal{H}(p_{\theta(\mu)}) & \text{if } \mu \in \mathcal{M}^\circ \\
+\infty & \text{if } \mu \notin \overline{\mathcal{M}} 
\end{cases}
$$

(b) Partition function has variational representation (dual of dual)

$$
A(\theta) = \sup_{\mu \in \mathcal{M}} \{ \langle \theta, \mu \rangle - A^*(\mu) \}
$$

(c) For $\theta \in \Omega$, sup occurs at $\mu \in \mathcal{M}^\circ$ of moment matching conditions

$$
\mu = \int_{\mathcal{D}_X} \phi(x)p_{\theta}(x)\nu(dx) = \mathbb{E}_\theta[\phi(X)] = \nabla A(\theta)
$$
Expectation Propagation: basic idea

- Came from a method called “assumed density filtering” (ADF).
Expectation Propagation: basic idea

- Came from a method called “assumed density filtering” (ADF).
- Doing full inference involves exponential computation.
Expectation Propagation: basic idea

- Came from a method called “assumed density filtering” (ADF).
- Doing full inference involves exponential computation.
- We do a bit of inference, involving reasonable computation, and getting us a new distribution that is a bit more complex but not too much more complex.
Expectation Propagation: basic idea

- Came from a method called “assumed density filtering” (ADF).
- Doing full inference involves exponential computation.
- We do a bit of inference, involving reasonable computation, and getting us a new distribution that is a bit more complex but not too much more complex.
- Before going further, we “project” this new distribution back down to a class of simple distributions.
Expectation Propagation: basic idea

- Came from a method called “assumed density filtering” (ADF).
- Doing full inference involves exponential computation.
- We do a bit of inference, involving reasonable computation, and getting us a new distribution that is a bit more complex but not too much more complex.
- Before going further, we “project” this new distribution back down to a class of simple distributions.
- We then repeat the above step with a bit more of inference, different than what we did above.
Expectation Propagation: basic idea

- Came from a method called “assumed density filtering” (ADF).
- Doing full inference involves exponential computation.
- We do a bit of inference, involving reasonable computation, and getting us a new distribution that is a bit more complex but not too much more complex.
- Before going further, we “project” this new distribution back down to a class of simple distributions.
- We then repeat the above step with a bit more of inference, different than what we did above.
- We keep repeating: do a bit of inference, and project, until all inference has been done.
Expectation Propagation: basic idea

- Came from a method called “assumed density filtering” (ADF).
- Doing full inference involves exponential computation.
- We do a bit of inference, involving reasonable computation, and getting us a new distribution that is a bit more complex but not too much more complex.
- Before going further, we “project” this new distribution back down to a class of simple distributions.
- We then repeat the above step with a bit more of inference, different than what we did above.
- We keep repeating: do a bit of inference, and project, until all inference has been done.
- The difference between ADF and EP is that, with ADF at this stage we’re done. With EP we can keep repeating the process of inference, projection.
Expectation Propagation: basic idea

- Came from a method called “assumed density filtering” (ADF).
- Doing full inference involves exponential computation.
- We do a bit of inference, involving reasonable computation, and getting us a new distribution that is a bit more complex but not too much more complex.
- Before going further, we “project” this new distribution back down to a class of simple distributions.
- We then repeat the above step with a bit more of inference, different than what we did above.
- We keep repeating: do a bit of inference, and project, until all inference has been done.
- The difference between ADF and EP is that, with ADF at this stage we’re done. With EP we can keep repeating the process of inference, projection.
- EP can be seen as a generalization of BP.
Expectation Propagation: basic idea

- Came from a method called “assumed density filtering” (ADF).
- Doing full inference involves exponential computation.
- We do a bit of inference, involving reasonable computation, and getting us a new distribution that is a bit more complex but not too much more complex.
- Before going further, we “project” this new distribution back down to a class of simple distributions.
- We then repeat the above step with a bit more of inference, different than what we did above.
- We keep repeating: do a bit of inference, and project, until all inference has been done.
- The difference between ADF and EP is that, with ADF at this stage we’re done. With EP we can keep repeating the process of inference, projection.
- EP can be seen as a generalization of BP.
- Interestingly, EP is instance of our variational framework, Equation ??.
Partition the $d$ sufficient statistics into two parts, the tractable ones (of which there are $d_T$) and the intractable ones (of which there are $d_I$). Thus, $d = d_T + d_I$. 

\[ \phi = \sum_{x \in \mathcal{X}} \log p(x) \]

\[ \phi_i \text{ are typically univariate, while } \Phi_i \text{ are multivariate (}b\text{-dimensional).} \]
Term Decoupling

- Partition the $d$ sufficient statistics into two parts, the tractable ones (of which there are $d_T$) and the intractable ones (of which there are $d_I$). Thus, $d = d_T + d_I$.

- Tractable component

$$\phi \triangleq (\phi_1, \phi_2, \ldots, \phi_{d_T}) \quad (16.5)$$
Term Decoupling

- Partition the $d$ sufficient statistics into two parts, the tractable ones (of which there are $d_T$) and the intractable ones (of which there are $d_I$). Thus, $d = d_T + d_I$.

- Tractable component

  \[ \phi \triangleq (\phi_1, \phi_2, \ldots, \phi_{d_T}) \]  
  \hfill (16.5)

- Intractable component

  \[ \Phi \triangleq (\Phi^1, \Phi^2, \ldots, \Phi^{d_I}) \]  
  \hfill (16.6)
Term Decoupling

- Partition the $d$ sufficient statistics into two parts, the tractable ones (of which there are $d_T$) and the intractable ones (of which there are $d_I$). Thus, $d = d_T + d_I$.

- Tractable component
  \[
  \phi \triangleq (\phi_1, \phi_2, \ldots, \phi_{d_T}) \tag{16.5}
  \]

- Intractable component
  \[
  \Phi \triangleq (\Phi^1, \Phi^2, \ldots, \Phi^{d_I}) \tag{16.6}
  \]

- $\phi_i$ are typically univariate, while $\Phi^i$ are multivariate ($b$-dimensional).
Term Decoupling

- Partition the $d$ sufficient statistics into two parts, the tractable ones (of which there are $d_T$) and the intractable ones (of which there are $d_I$). Thus, $d = d_T + d_I$.

- Tractable component

$$\phi \triangleq (\phi_1, \phi_2, \ldots, \phi_{d_T}) \quad (16.5)$$

- Intractable component

$$\Phi \triangleq (\Phi^1, \Phi^2, \ldots, \Phi^{d_I}) \quad (16.6)$$

- $\phi_i$ are typically univariate, while $\Phi^i$ are multivariate ($b$-dimensional).

- Consider exponential families associated with subcollection $(\phi, \Phi)$. 
Tractable component

\[ \phi \triangleq (\phi_1, \phi_2, \ldots, \phi_{d_T}) \]  \hfill (16.7)
Tractable component

\[ \phi \triangleq (\phi_1, \phi_2, \ldots, \phi_{dT}) \]  \hspace{1cm} (16.7)

So \( \phi : \mathcal{X}^m \rightarrow \mathbb{R}^{dT} \) with vector of parameters \( \theta \in \mathbb{R}^{dT} \).
Tractable component

\[ \phi \triangleq (\phi_1, \phi_2, \ldots, \phi_{d_T}) \quad (16.7) \]

- So \( \phi : \mathcal{X}^m \rightarrow \mathbb{R}^{d_T} \) with vector of parameters \( \theta \in \mathbb{R}^{d_T} \).
- Could instantiate model based only on this subcomponent, called the base model.
\[ \Phi \triangleq (\Phi_1, \Phi_2, \ldots, \Phi_{d_I}) \]
Intractable component

\[ \Phi \triangleq (\Phi_1, \Phi_2, \ldots, \Phi_{d_I}) \]  \hspace{1cm} (16.8)

Each \( \Phi_i : \mathcal{X}^m \rightarrow \mathbb{R}^b \).
Intractable component

\[ \Phi \triangleq (\Phi_1, \Phi_2, \ldots, \Phi_{d_I}) \] (16.8)

- Each \( \Phi_i : \mathcal{X}^m \rightarrow \mathbb{R}^b \).
- \( \Phi : \mathcal{X}^m \rightarrow \mathbb{R}^{b \times d_I} \).
Intractable component

\[ \Phi \triangleq (\Phi_1, \Phi_2, \ldots, \Phi_{d_I}) \quad (16.8) \]

- Each \( \Phi_i : \mathcal{X}^m \to \mathbb{R}^b \).
- \( \Phi : \mathcal{X}^m \to \mathbb{R}^{b \times d_I} \).
- Parameters \( \tilde{\theta} \in \mathbb{R}^{b \times d_I} \).
Associated Distributions

- The associated exponential family

\[
p(x; \theta, \tilde{\theta}) \propto \exp(\langle \theta, \phi(x) \rangle) \exp(\langle \tilde{\theta}, \Phi(x) \rangle)
\]

(16.9)

\[
= \exp(\langle \theta, \phi(x) \rangle) \prod_{i=1}^{d_I} \exp(\langle \tilde{\theta}_i, \Phi^i(x) \rangle)
\]

(16.10)
The associated exponential family

\[ p(x; \theta, \tilde{\theta}) \propto \exp (\langle \theta, \phi(x) \rangle) \exp (\langle \tilde{\theta}, \Phi(x) \rangle) \]  \hspace{1cm} (16.9)

\[ = \exp (\langle \theta, \phi(x) \rangle) \prod_{i=1}^{d_I} \exp (\langle \tilde{\theta}^i, \Phi^i(x) \rangle) \]  \hspace{1cm} (16.10)

Base model is tractable

\[ p(x; \theta, \tilde{0}) \propto \exp (\langle \theta, \phi(x) \rangle) \]  \hspace{1cm} (16.11)
The associated exponential family

\[
p(x; \theta, \tilde{\theta}) \propto \exp(\langle \theta, \phi(x) \rangle) \exp\left(\langle \tilde{\theta}, \Phi(x) \rangle\right) \tag{16.9}
\]

\[
= \exp(\langle \theta, \phi(x) \rangle) \prod_{i=1}^{d_I} \exp\left(\langle \tilde{\theta}^i, \Phi^i(x) \rangle\right) \tag{16.10}
\]

Base model is tractable

\[
p(x; \theta, \tilde{\theta}) \propto \exp(\langle \theta, \phi(x) \rangle) \tag{16.11}
\]

\[\Phi^i\text{-augmented model}\]

\[
p(x; \theta, \tilde{\theta}^i) \propto \exp(\langle \theta, \phi(x) \rangle) \exp\left(\langle \tilde{\theta}^i, \Phi^i(x) \rangle\right) \tag{16.12}
\]
The basic premises in the tractable-intractable partitioning between $\phi$ and $\Phi$ are:

- **It is possible to compute marginals exactly in polynomial time for distributions of the base form** (any member of the $\phi$-exponential family).

- Intractable to perform exact computations with the full $(\phi, \Phi)$-exponential family.
The basic premises in the tractable-intractable partitioning between \( \phi \) and \( \Phi \) are:

- It is possible to compute marginals exactly in polynomial time for distributions of the base form (any member of the \( \phi \)-exponential family).
- For each \( i = 1, \ldots, d_I \), exact polynomial-time computation is still possible for any \( \Phi^i \)-augmented form (any member of the \( (\phi, \Phi^i) \)-exponential family).
The basic premises in the tractable-intractable partitioning between $\phi$ and $\Phi$ are:

- It is possible to compute marginals exactly in polynomial time for distributions of the base form (any member of the $\phi$-exponential family).
- For each $i = 1, \ldots, d_I$, exact polynomial-time computation is still possible for any $\Phi^i$-augmented form (any member of the $(\phi, \Phi^i)$-exponential family).
- Intractable to perform exact computations with the full $(\phi, \Phi)$-exponential family.
Example: Mixture models

- Let $X \in \mathbb{R}^m$ be Gaussian with distribution $N(0, \Sigma)$.
Example: Mixture models

- Let $X \in \mathbb{R}^m$ be Gaussian with distribution $N(0, \Sigma)$.
- Let $\varphi(y; \mu, \Lambda)$ be Gaussian with mean $\mu$ covariance $\Lambda$. 

\[ p(y|X=x) = \begin{cases} (1-\alpha) \varphi(y; 0, \sigma_0^2 I) + \alpha \varphi(y; x, \sigma_1^2 I) \end{cases} \quad (16.13) \]
Example: Mixture models

- Let $X \in \mathbb{R}^m$ be Gaussian with distribution $N(0, \Sigma)$.
- Let $\varphi(y; \mu, \Lambda)$ be Gaussian with mean $\mu$ covariance $\Lambda$.
- Suppose $y$ conditioned on $x$ is a two-component Gaussian mixture model taking the form:

$$p(y|X = x) = (1 - \alpha)\varphi(y; 0, \sigma_0^2 I) + \alpha\varphi(y; x, \sigma_1^2 I)$$

(16.13)
Example: Mixture models

- Let $X \in \mathbb{R}^m$ be Gaussian with distribution $N(0, \Sigma)$.
- Let $\varphi(y; \mu, \Lambda)$ be Gaussian with mean $\mu$ covariance $\Lambda$.
- Suppose $y$ conditioned on $x$ is a two-component Gaussian mixture model taking the form:

$$p(y|X = x) = (1 - \alpha)\varphi(y; 0, \sigma_0^2 I) + \alpha\varphi(y; x, \sigma_1^2 I)$$

(16.13)

- Assume we have obtained $n$ i.i.d. samples $y^1, \ldots, y^n$ from mixture density, and goal is to produce posterior $p(x|y^1, \ldots, y^n)$, similar to Bayes-rule inverting a Naive-Bayes model.
Example: Mixture models

- Let $X \in \mathbb{R}^m$ be Gaussian with distribution $N(0, \Sigma)$.
- Let $\varphi(y; \mu, \Lambda)$ be Gaussian with mean $\mu$ covariance $\Lambda$.
- Suppose $y$ conditioned on $x$ is a two-component Gaussian mixture model taking the form:

$$p(y|X = x) = (1 - \alpha)\varphi(y; 0, \sigma_0^2 I) + \alpha\varphi(y; x, \sigma_1^2 I)$$  \hspace{1cm} (16.13)

- Assume we have obtained $n$ i.i.d. samples $y^1, \ldots, y^n$ from mixture density, and goal is to produce posterior $p(x|y^1, \ldots, y^n)$, similar to Bayes-rule inverting a Naive-Bayes model.

- Using Bayes rule, we get mixture model with $2^n$ components!

$$p(x|y^1, \ldots, y^n) \propto \exp \left( -\frac{1}{2} x^T \Sigma^{-1} x \right) \prod_{i=1}^n p(y^i|X = x)$$  \hspace{1cm} (16.14)

$$= \exp \left( -\frac{1}{2} x^T \Sigma^{-1} x \right) \exp \left\{ \sum_{i=1}^n \log p(y^i|X = x) \right\}$$  \hspace{1cm} (16.15)
Example: Mixture models

- We equate \( \exp\left(-\frac{1}{2}x^\top \sigma^{-1} x \right) \) with \( \exp(\langle \theta, \phi(x) \rangle) \), with \( d_T = m \).
Example: Mixture models

- We equate $\exp\left(-\frac{1}{2}x^\top \sigma^{-1} x\right)$ with $\exp(\langle \theta, \phi(x) \rangle)$, with $d_T = m$.
- Such a distribution is multivariate Gaussian, and getting marginals (say $p(x_A)$ for $A \subseteq [m]$) from it is relatively “cheap” $O(m^3)$.
Example: Mixture models

- We equate \( \exp \left( -\frac{1}{2} x^T \sigma^{-1} x \right) \) with \( \exp(\langle \theta, \phi(x) \rangle) \), with \( d_T = m \).
- Such a distribution is multivariate Gaussian, and getting marginals (say \( p(x_A) \) for \( A \subseteq [m] \)) from it is relatively “cheap” \( O(m^3) \).
- \( \exp \left\{ \sum_{i=1}^{n} \log p(y^i|X = x) \right\} \) equates to \( \prod_{i=1}^{d_I} \exp \left( \langle \tilde{\theta}^i, \Phi^i(x) \rangle \right) \), with \( b = 1 \). These are the intractable factors.
Example: Mixture models

- We equate $\exp\left(-\frac{1}{2}x^T \sigma^{-1} x\right)$ with $\exp(\langle \theta, \phi(x) \rangle)$, with $d_T = m$.
- Such a distribution is multivariate Gaussian, and getting marginals (say $p(x_A)$ for $A \subseteq [m]$) from it is relatively “cheap” $O(m^3)$.
- $\exp \left\{ \sum_{i=1}^{n} \log p(y^i | X = x) \right\}$ equates to $\prod_{i=1}^{d_I} \exp \left( \langle \tilde{\theta}^i, \Phi^i(x) \rangle \right)$, with $b = 1$. These are the intractable factors.
- Base distribution $p(x; \theta, \tilde{0}) \propto \exp \left(-\frac{1}{2}x^T \sigma^{-1} x\right)$ which is a Gaussian and easy as mentioned above.
Example: Mixture models

- We equate \( \exp \left( -\frac{1}{2} x^\top \sigma^{-1} x \right) \) with \( \exp(\langle \theta, \phi(x) \rangle) \), with \( d_T = m \).
- Such a distribution is multivariate Gaussian, and getting marginals (say \( p(x_A) \) for \( A \subseteq [m] \)) from it is relatively “cheap” \( O(m^3) \).
- \( \exp \left\{ \sum_{i=1}^{n} \log p(y^i|X=x) \right\} \) equates to \( \prod_{i=1}^{d_I} \exp \left( \langle \tilde{\theta}^i, \Phi^i(x) \rangle \right) \), with \( b = 1 \). These are the intractable factors.
- Base distribution \( p(x; \theta, \vec{0}) \propto \exp \left( -\frac{1}{2} x^\top \sigma^{-1} x \right) \) which is a Gaussian and easy as mentioned above.
- If we multiply in only one intractable term, complexity to produce marginal still not so bad (quite easy in fact).
Example: Mixture models

- We equate $\exp \left( -\frac{1}{2}x^\top \sigma^{-1}x \right)$ with $\exp(\langle \theta, \phi(x) \rangle)$, with $d_T = m$.
- Such a distribution is multivariate Gaussian, and getting marginals (say $p(x_A)$ for $A \subseteq [m]$) from it is relatively “cheap” $O(m^3)$.
- $\exp \left\{ \sum_{i=1}^{n} \log p(y^i|X = x) \right\}$ equates to $\prod_{i=1}^{d_I} \exp \left( \langle \tilde{\theta}^i, \Phi^i(x) \rangle \right)$, with $b = 1$. These are the intractable factors.
- Base distribution $p(x; \theta, \vec{0}) \propto \exp \left( -\frac{1}{2}x^\top \sigma^{-1}x \right)$ which is a Gaussian and easy as mentioned above.
- If we multiply in only one intractable term, complexity to produce marginal still not so bad (quite easy in fact).
- I.e., $\Phi^i$-augmented distribution is proportional to

$$\exp \left( -\frac{1}{2}x^\top \sigma^{-1}x \right) \left[ (1 - \alpha)\varphi(y^i; 0, \sigma_0^2 I) + \alpha \varphi(y^i; x, \sigma_1^2 I) \right] \quad (16.16)$$
Example: Mixture models

- We equate $\exp \left( -\frac{1}{2} x^\top \sigma^{-1} x \right)$ with $\exp(\langle \theta, \phi(x) \rangle)$, with $d_T = m$.
- Such a distribution is multivariate Gaussian, and getting marginals (say $p(x_A)$ for $A \subseteq [m]$) from it is relatively “cheap” $O(m^3)$.
- $\exp \left\{ \sum_{i=1}^{n} \log p(y^i | X = x) \right\}$ equates to $\prod_{i=1}^{d_I} \exp \left( \langle \tilde{\theta}^i, \Phi^i(x) \rangle \right)$, with $b = 1$. These are the intractable factors.
- Base distribution $p(x; \theta, \vec{0}) \propto \exp \left( -\frac{1}{2} x^\top \sigma^{-1} x \right)$ which is a Gaussian and easy as mentioned above.
- If we multiply in only one intractable term, complexity to produce marginal still not so bad (quite easy in fact).
- I.e., $\Phi^i$-augmented distribution is proportional to

$$\exp \left( -\frac{1}{2} x^\top \sigma^{-1} x \right) \left[ (1 - \alpha) \phi(y^i; 0, \sigma^2_0 I) + \alpha \phi(y^i; x, \sigma^2_1 I) \right] \quad (16.16)$$

- Computing marginals is easy (mixture of only 2 components)
Example: Mixture models

- We equate \( \exp\left(-\frac{1}{2}x^\top \sigma^{-1} x\right) \) with \( \exp(\langle \theta, \phi(x) \rangle) \), with \( d_T = m \).
- Such a distribution is multivariate Gaussian, and getting marginals (say \( p(x_A) \) for \( A \subseteq [m] \)) from it is relatively “cheap” \( O(m^3) \).
- \( \exp \left\{ \sum_{i=1}^{n} \log p(y^i|X = x) \right\} \) equates to \( \prod_{i=1}^{d_f} \exp \left( \langle \tilde{\theta}^i, \Phi^i(x) \rangle \right) \), with \( b = 1 \). These are the intractable factors.
- Base distribution \( p(x; \theta, \vec{0}) \propto \exp\left(-\frac{1}{2}x^\top \sigma^{-1} x\right) \) which is a Gaussian and easy as mentioned above.
- If we multiply in only one intractable term, complexity to produce marginal still not so bad (quite easy in fact).
- I.e., \( \Phi^i \)-augmented distribution is proportional to

\[
\exp \left( -\frac{1}{2} x^\top \sigma^{-1} x \right) \left[ (1 - \alpha) \varphi(y^i; 0, \sigma_0^2 I) + \alpha \varphi(y^i; x, \sigma_1^2 I) \right] \quad (16.16)
\]

- Computing marginals is easy (mixture of only 2 components)
- If we multiply in all \( \Phi^i \), becomes intractable (\( 2^m \) potentially distinct components each of which requires marginalization).
Polytope and Base case

- We can partition the mean parameters $(\mu, \tilde{\mu}) \in \mathbb{R}^{d_T+d_I \times b}$.
We can partition the mean parameters \((\mu, \tilde{\mu}) \in \mathbb{R}^{dT + dI \times b}\)

Marginal polytope associated with these means

\[
\mathcal{M}(\phi, \Phi) = \{(\mu, \tilde{\mu})| (\mu, \tilde{\mu}) = \mathbb{E}_p[(\phi(X), \Phi(X))] \text{ for some } p\}
\]  (16.17)

along with negative dual of cumulant, or entropy

\[
H(\mu, \tilde{\mu}) = -A^*(\mu, \tilde{\mu}).
\]
Polytope and Base case

- We can partition the mean parameters \((\mu, \tilde{\mu}) \in \mathbb{R}^{d_T+d_I \times b}\)
- Marginal polytope associated with these means

\[
\mathcal{M}(\phi, \Phi) = \{(\mu, \tilde{\mu})| (\mu, \tilde{\mu}) = \mathbb{E}_p[(\phi(X), \Phi(X))]| \text{ for some } p\} \tag{16.17}
\]

along with negative dual of cumulant, or entropy

\[
H(\mu, \tilde{\mu}) = -A^*(\mu, \tilde{\mu}).
\]

- We also have polytope associated with only base distribution

\[
\mathcal{M}(\phi) = \left\{ \mu \in \mathbb{R}^{d_T} | \mu = \mathbb{E}_p(\phi(X)) \right\} \tag{16.18}
\]
We can partition the mean parameters \((\mu, \tilde{\mu}) \in \mathbb{R}^{d_T+d_I \times b}\)

Marginal polytope associated with these means

\[
\mathcal{M}(\phi, \Phi) = \{ (\mu, \tilde{\mu}) | (\mu, \tilde{\mu}) = \mathbb{E}_p[(\phi(X), \Phi(X))] \text{ for some } p \} \quad (16.17)
\]

along with negative dual of cumulant, or entropy

\[
H(\mu, \tilde{\mu}) = -A^*(\mu, \tilde{\mu}).
\]

We also have polytope associated with only base distribution

\[
\mathcal{M}(\phi) = \{ \mu \in \mathbb{R}^{d_T} | \mu = \mathbb{E}_p(\phi(X)) \} \quad (16.18)
\]

Recall thm: any mean in the interior is realizable via an exponential family model, and associated entropy \(H(\mu)\) is tractable.
Augmented Base case

For each $i = 1 \ldots d_I$ we have a $\Phi^i$-augmented exp. model and polytope

$$\mathcal{M}(\phi, \Phi^i) = \left\{ (\mu, \tilde{\mu}^i) \in \mathbb{R}^{d_I + b} | (\mu, \tilde{\mu}^i) = \mathbb{E}_p[(\phi(X), \Phi^i(X))] \text{ for some } p \right\}$$

(16.19)
Augmented Base case

- For each $i = 1 \ldots d_I$ we have a $\Phi^i$-augmented exp. model and polytope

$$\mathcal{M}(\phi, \Phi^i) = \{(\mu, \tilde{\mu}^i) \in \mathbb{R}^{d_T + b} | (\mu, \tilde{\mu}^i) = \mathbb{E}_p[(\phi(X), \Phi^i(X))] \text{ for some } p \} \quad (16.19)$$

- Thus, any such mean parameters has instance for associated exponential family, and also $H(\mu, \tilde{\mu}^i)$ is easy to compute.
Augmented Base case

For each $i = 1 \ldots d_I$ we have a $\Phi^i$-augmented exp. model and polytope

$$M(\phi, \Phi^i) = \left\{ (\mu, \tilde{\mu}^i) \in \mathbb{R}^{d_I+b} | (\mu, \tilde{\mu}^i) = \mathbb{E}_p[(\phi(X), \Phi^i(X))] \right\} \text{ for some } p$$

(16.19)

Thus, any such mean parameters has instance for associated exponential family, and also $H(\mu, \tilde{\mu}^i)$ is easy to compute.

Goal, variational approximation: Need outer bounds on $M(\phi, \Phi)$ and expression for entropy (as is now normal).
Augmented Base case

- For each $i = 1 \ldots d_I$ we have a $\Phi^i$-augmented exp. model and polytope

$$\mathcal{M}(\phi, \Phi^i) = \left\{ (\mu, \tilde{\mu}^i) \in \mathbb{R}^{d_T+b} \middle| (\mu, \tilde{\mu}^i) = \mathbb{E}_p[(\phi(X), \Phi^i(X))] \text{ for some } p \right\}$$

(16.19)

- Thus, any such mean parameters has instance for associated exponential family, and also $H(\mu, \tilde{\mu}^i)$ is easy to compute.

- Goal, variational approximation: Need outer bounds on $\mathcal{M}(\phi, \Phi)$ and expression for entropy (as is now normal).

- Turns out we can do this, and an iterative algorithm to find fixed points of associated Lagrangian, that correspond to EP.
New outer bound

For any mean parms \((\tau, \tilde{\tau})\) where \(\tilde{\tau} = (\tilde{\tau}^1, \tilde{\tau}^2, \ldots, \tilde{\tau}^{d_I})\), define coordinate “projection operation”

\[
\Pi^i(\tau, \tilde{\tau}) \rightarrow (\tau, \tilde{\tau}^i)
\]  

(16.20)

This operator simply removes all but \(\tilde{\tau}^i\) from \(\tilde{\tau}\).
New outer bound

- For any mean parms \((\tau, \tilde{\tau})\) where \(\tilde{\tau} = (\tilde{\tau}^1, \tilde{\tau}^2, \ldots, \tilde{\tau}^{d_I})\), define coordinate "projection operation"

\[ \Pi^i(\tau, \tilde{\tau}) \rightarrow (\tau, \tilde{\tau}^i) \]  (16.20)

This operator simply removes all but \(\tilde{\tau}^i\) from \(\tilde{\tau}\).

- Define outer bound on true means \(M(\phi, \Phi)\) (which is still convex)

\[ \mathcal{L}(\phi, \Phi) = \{ (\tau, \tilde{\tau}) | \tau \in M(\phi), \Pi^i(\tau, \tilde{\tau}) \in M(\phi, \Phi^i), \forall i \} \]  (16.21)
New outer bound

- For any mean parms \((\tau, \tilde{\tau})\) where \(\tilde{\tau} = (\tilde{\tau}^1, \tilde{\tau}^2, \ldots, \tilde{\tau}^{d_I})\), define coordinate “projection operation”

\[
\Pi^i(\tau, \tilde{\tau}) \rightarrow (\tau, \tilde{\tau}^i)
\]  

(16.20)

This operator simply removes all but \(\tilde{\tau}^i\) from \(\tilde{\tau}\).

- Define outer bound on true means \(M(\phi, \Phi)\) (which is still convex)

\[
L(\phi, \Phi) = \{(\tau, \tilde{\tau}) | \tau \in M(\phi), \Pi^i(\tau, \tilde{\tau}) \in M(\phi, \Phi^i), \forall i\}
\]  

(16.21)

- Note, based on a set of projections onto \(M(\phi, \Phi^i)\). Clearly outer bound since \(M(\phi, \Phi) \subseteq L(\phi, \Phi)\).
New outer bound

- For any mean parms \((\tau, \tilde{\tau})\) where \(\tilde{\tau} = (\tilde{\tau}_1, \tilde{\tau}_2, \ldots, \tilde{\tau}_d)\), define coordinate “projection operation”

\[
\Pi^i(\tau, \tilde{\tau}) \rightarrow (\tau, \tilde{\tau}_i)
\]  \hspace{1cm} (16.20)

This operator simply removes all but \(\tilde{\tau}_i\) from \(\tilde{\tau}\).

- Define outer bound on true means \(M(\phi, \Phi)\) (which is still convex)

\[
\mathcal{L}(\phi, \Phi) = \{(\tau, \tilde{\tau}) | \tau \in M(\phi), \Pi^i(\tau, \tilde{\tau}) \in M(\phi, \Phi^i), \forall i\} \]  \hspace{1cm} (16.21)

- Note, based on a set of projections onto \(M(\phi, \Phi^i)\). Clearly outer bound since \(M(\phi, \Phi) \subseteq \mathcal{L}(\phi, \Phi)\).

- If \(\Phi^i\) are edges of a graph (i.e. local consistency) then we get standard \(\mathbb{L}\) outer bound we saw before with Bethe approximation.
Members in new outer bound

- For any mean parms \((\tau, \tilde{\tau}) \in \mathcal{L}(\phi, \Phi)\):
Members in new outer bound

- For any mean parms \((\tau, \tilde{\tau}) \in \mathcal{L}(\phi, \Phi)\): 
  A) There is a member of the \(\phi\)-exponential family which mean parameters \(\tau\) with entropy \(H(\tau)\);
For any mean params \((\tau, \tilde{\tau}) \in \mathcal{L} (\phi, \Phi)\): A) There is a member of the \(\phi\)-exponential family which mean parameters \(\tau\) with entropy \(H(\tau)\); B) Also, for \(i = 1 \ldots d_I\), there is a member of the \((\phi, \Phi^i)\)-exponential family with mean parameters \((\tau, \tilde{\tau}^i)\) with entropy \(H(\tau, \tilde{\tau}^i)\).
Members in new outer bound

- For any mean parms \((\tau, \tilde{\tau}) \in \mathcal{L}(\phi, \Phi)\): A) There is a member of the \(\phi\)-exponential family which mean parameters \(\tau\) with entropy \(H(\tau)\); B) Also, for \(i = 1 \ldots d_I\), there is a member of the \((\phi, \Phi^i)\)-exponential family with mean parameters \((\tau, \tilde{\tau}^i)\) with entropy \(H(\tau, \tilde{\tau}^i)\).

- Both entropy forms are easy to compute, and so is a new entropy approximation:

\[
H(\tau, \tilde{\tau}) \approx H_{ep}(\tau, \tilde{\tau}) \triangleq H(\tau) + \sum_{\ell=1}^{d_I} \left[ H(\tau, \tilde{\tau}^\ell) - H(\tau) \right]
\]

(16.22)
Members in new outer bound

- For any mean parms \((\tau, \tilde{\tau}) \in \mathcal{L}(\phi, \Phi)\): A) There is a member of the \(\phi\)-exponential family which mean parameters \(\tau\) with entropy \(H(\tau)\); B) Also, for \(i = 1 \ldots d_I\), there is a member of the \((\phi, \Phi^i)\)-exponential family with mean parameters \((\tau, \tilde{\tau}^i)\) with entropy \(H(\tau, \tilde{\tau}^i)\).

- Both entropy forms are easy to compute, and so is a new entropy approximation:

\[
H(\tau, \tilde{\tau}) \approx H_{ep}(\tau, \tilde{\tau}) \triangleq H(\tau) + \sum_{\ell=1}^{d_I} \left[ H(\tau, \tilde{\tau}^l) - H(\tau) \right] \quad (16.22)
\]

- With outer bound and entropy expression, we get new variational form:

\[
\max_{(\tau, \tilde{\tau}) \in \mathcal{L}(\phi, \Phi)} \left\{ \langle \tau, \theta \rangle + \langle \tilde{\tau}, \tilde{\theta} \rangle + H_{ep}(\tau, \tilde{\tau}) \right\} \quad (16.23)
\]
Members in new outer bound

For any mean parms \((\tau, \tilde{\tau}) \in \mathcal{L}(\phi, \Phi)\): A) There is a member of the \(\phi\)-exponential family which mean parameters \(\tau\) with entropy \(H(\tau)\); B) Also, for \(i = 1 \ldots d_I\), there is a member of the \((\phi, \Phi^i)\)-exponential family with mean parameters \((\tau, \tilde{\tau}^i)\) with entropy \(H(\tau, \tilde{\tau}^i)\).

Both entropy forms are easy to compute, and so is a new entropy approximation:

\[
H(\tau, \tilde{\tau}) \approx H_{ep}(\tau, \tilde{\tau}) \triangleq H(\tau) + \sum_{\ell=1}^{d_I} \left[ H(\tau, \tilde{\tau}^\ell) - H(\tau) \right] 
\]

(16.22)

With outer bound and entropy expression, we get new variational form

\[
\max_{(\tau, \tilde{\tau}) \in \mathcal{L}(\phi, \Phi)} \left\{ \langle \tau, \theta \rangle + \langle \tilde{\tau}, \tilde{\theta} \rangle + H_{ep}(\tau, \tilde{\tau}) \right\} 
\]

(16.23)

This characterizes the EP algorithms.
Members in new outer bound

- For any mean parms \((\tau, \tilde{\tau}) \in \mathcal{L}(\phi, \Phi)\): A) There is a member of the \(\phi\)-exponential family which mean parameters \(\tau\) with entropy \(H(\tau)\); B) Also, for \(i = 1 \ldots d_I\), there is a member of the \((\phi, \Phi^i)\)-exponential family with mean parameters \((\tau, \tilde{\tau}^i)\) with entropy \(H(\tau, \tilde{\tau}^i)\).

- Both entropy forms are easy to compute, and so is a new entropy approximation:

\[
H(\tau, \tilde{\tau}) \approx H_{ep}(\tau, \tilde{\tau}) \overset{\Delta}{=} H(\tau) + \sum_{\ell=1}^{d_I} \left[ H(\tau, \tilde{\tau}^l) - H(\tau) \right] \tag{16.22}
\]

- With outer bound and entropy expression, we get new variational form

\[
\max_{(\tau, \tilde{\tau}) \in \mathcal{L}(\phi, \Phi)} \left\{ \langle \tau, \theta \rangle + \langle \tilde{\tau}, \tilde{\theta} \rangle + H_{ep}(\tau, \tilde{\tau}) \right\} \tag{16.23}
\]

- This characterizes the EP algorithms.

- Given graph \(G = (V, E)\) when we take \(\phi\) to be unaries \(V\) and \(\Phi\) to be edges \(E\), we exactly recover Bethe approximation.
Lagrangian optimization setup

- Make $d_I$ duplicates of vector $\tau \in \mathbb{R}^{d_T}$, call them $\eta^i \in \mathbb{R}^{d_T}$ for $i \in [d_T]$. 

This gives large set of pseudo-mean parameters 

$$\{\tau, (\eta^i, \tilde{\tau}^i), i \in [d_I]\} \in \mathbb{R}^{d_T} \times (\mathbb{R}^{d_T} \times \mathbb{R}^{b})^{d_I}$$

We arrive at the optimization:

$$\max \{\tau, \{ (\eta^i, \tilde{\tau}^i) \}_{i \in [d_I]} \} \{ \langle \tau, \theta \rangle + d_I \sum_{i=1}^{d_I} \langle \tilde{\tau}^i, \tilde{\theta}^i \rangle + H(\tau) + d_I \sum_{i=1}^{d_I} [H(\eta^i, \tilde{\tau}^i) - H(\eta^i)] \}$$

subject to $\tau \in M(\phi)$, and for all $i$ that $\tau = \eta^i$ and that $(\eta^i, \tilde{\tau}^i) \in M(\phi, \Phi_i)$. 

Use Lagrange multipliers to impose constraint $\eta^i = \tau$ for all $i$, and for the rest of the constraints too.
Lagrangian optimization setup

- Make $d_I$ duplicates of vector $\tau \in \mathbb{R}^{d_T}$, call them $\eta^i \in \mathbb{R}^{d_T}$ for $i \in [d_T]$.
- This gives large set of pseudo-mean parameters

$$
\{ \tau, (\eta^i, \tilde{\tau}^i), i \in [d_I] \} \in \mathbb{R}^{d_T} \times (\mathbb{R}^{d_T} \times \mathbb{R}^b)^{d_I}
$$

(16.24)
Lagrangian optimization setup

- Make $d_I$ duplicates of vector $\tau \in \mathbb{R}^{dT}$, call them $\eta^i \in \mathbb{R}^{dT}$ for $i \in [dT]$.
- This gives large set of pseudo-mean parameters

$$\{\tau, (\eta^i, \tilde{\tau}^i), i \in [d_I]\} \in \mathbb{R}^{dT} \times (\mathbb{R}^{dT} \times \mathbb{R}^b)^{d_I} \quad (16.24)$$

- We arrive at the optimization:

$$\max \left\{ \langle \tau, \theta \rangle + \sum_{i=1}^{d_I} \langle \tilde{\tau}^i, \tilde{\theta}^i \rangle + H(\tau) + \sum_{i=1}^{d_I} [H(\eta^i, \tilde{\tau}^i) - H(\eta^i)] \right\} \quad (16.25)$$

subject to $\tau \in \mathcal{M}(\phi)$, and for all $i$ that $\tau = \eta^i$ and that $(\eta^i, \tilde{\tau}^i) \in \mathcal{M}(\phi, \Phi^i)$. 
Lagrangian optimization setup

- Make $d_I$ duplicates of vector $\tau \in \mathbb{R}^{dT}$, call them $\eta^i \in \mathbb{R}^{dT}$ for $i \in [dT]$.
- This gives large set of pseudo-mean parameters
  \[
  \{\tau, (\eta^i, \tilde{\tau}^i), i \in [d_I]\} \in \mathbb{R}^{dT} \times (\mathbb{R}^{dT} \times \mathbb{R}^b)^{d_I}
  \]  
  (16.24)
- We arrive at the optimization:

  \[
  \max \left\{ \langle \tau, \theta \rangle + \sum_{i=1}^{d_I} \langle \tilde{\tau}^i, \tilde{\theta}^i \rangle + H(\tau) + \sum_{i=1}^{d_I} [H(\eta^i, \tilde{\tau}^i) - H(\eta^i)] \right\}
  \]
  (16.25)

  subject to $\tau \in \mathcal{M}(\phi)$, and for all $i$ that $\tau = \eta^i$ and that $(\eta^i, \tilde{\tau}^i) \in \mathcal{M}(\phi, \Phi^i)$.
- Use Lagrange multipliers to impose constraint $\eta^i = \tau$ for all $i$, and for the rest of the constraints too.
To Lagrangian optimization

- We get a Lagrangian version of the objective

\[
L(\tau; \lambda) = \langle \tau, \theta \rangle + \sum_{i=1}^{d_I} \langle \tilde{\tau}^i, \tilde{\theta}^i \rangle + F(\tau; (\eta^i, \tilde{\tau}^i)) + \sum_{i=1}^{d_I} \langle \lambda^i, \tau - \eta^i \rangle + \ldots
\]

(16.26)

where

\[
F(\tau; (\eta^i, \tilde{\tau}^i)) = H(\tau) + \sum_{i=1}^{d_I} \left[ H(\eta^i, \tilde{\tau}^i) - H(\eta^i) \right]
\]

(16.27)

and where \( \lambda^i \) are the Lagrange multipliers associated with the constraint \( \eta^i = \tau \) for all \( i \) (other multipliers not shown).
To Lagrangian optimization to Moment Matching

Considering optimality conditions on what must hold for a solution \( \{ \tau, (\eta^i, \tilde{\tau}^i), i \in [d_I] \} \) to the above Lagrangian, must have properties:
To Lagrangian optimization to Moment Matching

Considering optimality conditions on what must hold for a solution \( \{\tau, (\eta^i, \tilde{\tau}^i), i \in [d_I]\} \) to the above Lagrangian, must have properties:

1. \( \tau \) belongs to relative interior, i.e., \( \tau \in M^\circ(\theta) \) of the base model.
To Lagrangian optimization to Moment Matching

- Considering optimality conditions on what must hold for a solution \( \{\tau, (\eta^i, \tilde{\tau}^i), i \in [d_I] \} \) to the above Lagrangian, must have properties:
  1. \( \tau \) belongs to relative interior, i.e., \( \tau \in M^\circ(\theta) \) of the base model.
  2. \( (\eta^i, \tilde{\tau}^i) \) belongs to relative interior of extended model, so
     \( (\eta^i, \tilde{\tau}^i) \in M^\circ(\phi, \Phi^i) \).
To Lagrangian optimization to Moment Matching

Considering optimality conditions on what must hold for a solution \( \{\tau, (\eta^i, \tilde{\tau}^i), i \in [d_I]\} \) to the above Lagrangian, must have properties:

1. \( \tau \) belongs to relative interior, i.e., \( \tau \in \mathcal{M}^\circ(\theta) \) of the base model.
2. \( (\eta^i, \tilde{\tau}^i) \) belongs to relative interior of extended model, so \( (\eta^i, \tilde{\tau}^i) \in \mathcal{M}^\circ(\phi, \Phi^i) \).
3. Means must agree, i.e., \( \tau = \eta^i \) for all \( i \).
To Lagrangian optimization to Moment Matching

- Considering optimality conditions on what must hold for a solution \( \{ \tau, (\eta^i, \tilde{\tau}^i), i \in [d_I] \} \) to the above Lagrangian, must have properties:
  1. \( \tau \) belongs to relative interior, i.e., \( \tau \in M^\circ(\theta) \) of the base model.
  2. \((\eta^i, \tilde{\tau}^i)\) belongs to relative interior of extended model, so \((\eta^i, \tilde{\tau}^i) \in M^\circ(\phi, \Phi^i)\).
  3. Means must agree, i.e., \( \tau = \eta^i \) for all \( i \).

- First condition means we're a member of the \( \phi \)-exponential family, and (it can be shown) has form:

\[
q(x; \theta, \lambda) \propto \exp \left\{ \langle \theta + \sum_{i=1}^{d_I} \lambda^i, \phi(x) \rangle \right\} \tag{16.28}
\]
To Lagrangian optimization to Moment Matching

- Considering optimality conditions on what must hold for a solution \( \{\tau, (\eta^i, \tilde{\tau}^i), i \in [d_I]\} \) to the above Lagrangian, must have properties:
  1. \( \tau \) belongs to relative interior, i.e., \( \tau \in M^\circ(\theta) \) of the base model.
  2. \((\eta^i, \tilde{\tau}^i)\) belongs to relative interior of extended model, so \((\eta^i, \tilde{\tau}^i) \in M^\circ(\phi, \Phi^i)\).
  3. Means must agree, i.e., \( \tau = \eta^i \) for all \( i \).

- First condition means we’re a member of the \( \phi \)-exponential family, and (it can be shown) has form:

\[
q(x; \theta, \lambda) \propto \exp \left\{ \left\langle \theta + \sum_{i=1}^{d_I} \lambda^i, \phi(x) \right\rangle \right\}
\] (16.28)

- Second condition means we’re a member of the \( (\phi, \Phi^i) \)-exponential family, and (it can be shown) has form:

\[
q^i(x, \theta, \tilde{\theta}^i, \lambda) \propto \exp \left( \left\langle \theta + \sum_{\ell \neq i} \lambda^\ell, \phi(x) \right\rangle + \left\langle \tilde{\theta}^i, \Phi^i(x) \right\rangle \right) \] (16.29)
This condition is a form of moment-matching. I.e., we have $\tau = E_q[\phi(X)]$ and $\eta^i = E_{q^i}[\phi(X)]$, so equating these gives:

$$\int q(x; \theta, \lambda) \phi(x) \nu(dx) = \int q^i(x; \theta, \tilde{\theta}^i) \phi(x) \nu(dx)$$

(16.30)

for $i \in [d_I]$. 
1. At iteration $n = 0$, initialize the Lagrange multiplier vectors $(\lambda^1, \ldots, \lambda^{d_I})$.
At iteration $n = 0$, initialize the Lagrange multiplier vectors $(\lambda^1, \ldots, \lambda^{d_I})$

At each iteration $n = 1, 2, \ldots$ choose some index $i(n) \in \{1, \ldots, d_I\}$. 
1. At iteration $n = 0$, initialize the Lagrange multiplier vectors $(\lambda^1, \ldots, \lambda^{d_I})$.

2. At each iteration $n = 1, 2, \ldots$ choose some index $i(n) \in \{1, \ldots, d_I\}$.

3. Under the following augmented distribution

$$ q^i(x; \theta, \tilde{\theta}^i, \lambda) \propto \exp \left( \left\langle \theta + \sum_{\ell \neq i} \lambda^\ell, \phi(x) \right\rangle + \left\langle \tilde{\theta}^i, \Phi^i(x) \right\rangle \right), \quad (16.31) $$

compute the mean parameters $\eta^i$ as follows:

$$ \eta^{i(n)} = \int q^{i(n)}(x) \phi(x) \nu(dx) = \mathbb{E}_{q^{i(n)}}[\phi(X)] \quad (16.32) $$
Moment Matching $\rightarrow$ Expectation Propagation Updates

1. At iteration $n = 0$, initialize the Lagrange multiplier vectors $(\lambda^1, \ldots, \lambda^{d_I})$

2. At each iteration $n = 1, 2, \ldots$ choose some index $i(n) \in \{1, \ldots, d_I\}$.

3. Under the following augmented distribution

$$ q^i(x; \theta, \tilde{\theta}^i, \lambda) \propto \exp \left( \langle \theta + \sum_{\ell \neq i} \lambda^\ell, \phi(x) \rangle + \langle \tilde{\theta}^i, \Phi^i(x) \rangle \right) , \quad (16.31) $$

compute the mean parameters $\eta^i$ as follows:

$$ \eta^{i(n)} = \int q^{i(n)}(x) \phi(x) \nu(dx) = \mathbb{E}_{q^{i(n)}}[\phi(X)] \quad (16.32) $$

4. Form base distribution $q$ using Equation 16.28 and adjust $\lambda^{i(n)}$ to satisfy the moment-matching condition

$$ \mathbb{E}_q[\phi(X)] = \eta^{i(n)} \quad (16.33) $$
Moment Matching → Expectation Propagation Updates

1. At iteration $n = 0$, initialize the Lagrange multiplier vectors $(\lambda^1, \ldots, \lambda^{d_I})$.
2. At each iteration $n = 1, 2, \ldots$ choose some index $i(n) \in \{1, \ldots, d_I\}$.
3. Under the following augmented distribution

$$q^i(x; \theta, \tilde{\theta}^i, \lambda) \propto \exp \left( \left\langle \theta + \sum_{\ell \neq i} \lambda^\ell, \phi(x) \right\rangle + \left\langle \tilde{\theta}^i, \Phi^i(x) \right\rangle \right), \quad (16.31)$$

compute the mean parameters $\eta^i$ as follows:

$$\eta^i(n) = \int q^{i(n)}(x) \phi(x) \nu(dx) = \mathbb{E}_{q^{i(n)}}[\phi(X)] \quad (16.32)$$

4. Form base distribution $q$ using Equation 16.28 and adjust $\lambda^{i(n)}$ to satisfy the moment-matching condition

$$\mathbb{E}_q[\phi(X)] = \eta^i(n) \quad (16.33)$$

5. This is a KL-divergence minimization step, but done w. exponential family models which thus corresponds to moment-matching.
Example: Tree-structured EP

- When base distribution is a tree, we get what is called tree-structured EP
Example: Tree-structured EP

- When base distribution is a tree, we get what is called tree-structured EP
- Start with a graph $G = (V, E)$ and form a spanning tree $T = (V, E(T))$ in any arbitrary way.
Example: Tree-structured EP

- When base distribution is a tree, we get what is called tree-structured EP
- Start with a graph $G = (V, E)$ and form a spanning tree $T = (V, E(T))$ in any arbitrary way.
- Form base distribution as follows:

$$p(x; \theta, \vec{0}) \propto \prod_{s \in V} \exp(\theta_s(x_s)) \prod_{(s,t) \in E(T)} \exp(\theta_{st}(x_s, x_t))$$  \hspace{1cm} (16.34)
Example: Tree-structured EP

- When base distribution is a tree, we get what is called tree-structured EP.

- Start with a graph $G = (V, E)$ and form a spanning tree $T = (V, E(T))$ in any arbitrary way.

- Form base distribution as follows:

$$p(x; \theta, \vec{0}) \propto \prod_{s \in V} \exp(\theta_s(x_s)) \prod_{(s,t) \in E(T)} \exp(\theta_{st}(x_s, x_t))$$  \hspace{1cm} (16.34)

- Then, each $\Phi_i$ corresponds to an edge in $E \setminus E(T)$, and gives us, for each edge $(u, v) \in E \setminus E(T)$, the $\phi^{(u,v)}$-augmented distribution

$$p(x; \theta, \theta_{u,v}) \propto (x; \theta, \vec{0}) \exp(\theta_{u,v}(x_u, x_v))$$  \hspace{1cm} (16.35)
EP as variational: Summary of key points

- Fixed points of EP exist assuming Lagrangian form has at least one optimum.
EP as variational: Summary of key points

- Fixed points of EP exist assuming Lagrangian form has at least one optimum.
- No guarantees that EP will converge, but if it does it will be at a stationary point of the Lagrangian.
EP as variational: Summary of key points

- Fixed points of EP exist assuming Lagrangian form has at least one optimum.
- No guarantees that EP will converge, but if it does it will be at a stationary point of the Lagrangian.
- EP can be seen to be based on variational framework, using Bethe-like entropy and convex outer bound for the mean parameters.
EP as variational: Summary of key points

- Fixed points of EP exist assuming Lagrangian form has at least one optimum.
- No guarantees that EP will converge, but if it does it will be at a stationary point of the Lagrangian.
- EP can be seen to be based on variational framework, using Bethe-like entropy and convex outer bound for the mean parameters.
- When base distribution is unaries and $\Phi^i$ is the edges of a graph, we in fact get standard Bethe approximation, and standard sum-product LBP.
EP as variational: Summary of key points

- Fixed points of EP exist assuming Lagrangian form has at least one optimum.
- No guarantees that EP will converge, but if it does it will be at a stationary point of the Lagrangian.
- EP can be seen to be based on variational framework, using Bethe-like entropy and convex outer bound for the mean parameters.
- When base distribution is unaries and $\Phi^i$ is the edges of a graph, we in fact get standard Bethe approximation, and standard sum-product LBP.
- Moment matching of EP can be seen as striving for solution of associated Lagrangian.
EP as variational: Summary of key points

- Fixed points of EP exist assuming Lagrangian form has at least one optimum.
- No guarantees that EP will converge, but if it does it will be at a stationary point of the Lagrangian.
- EP can be seen to be based on variational framework, using Bethe-like entropy and convex outer bound for the mean parameters.
- When base distribution is unaries and $\Phi^i$ is the edges of a graph, we in fact get standard Bethe approximation, and standard sum-product LBP.
- Moment matching of EP can be seen as striving for solution of associated Lagrangian.
- Lost of flexibility here, depending on what the base distribution is (e.g., could be a $k$-tree or any other structure).
Fixed points of EP exist assuming Lagrangian form has at least one optimum.

No guarantees that EP will converge, but if it does it will be at a stationary point of the Lagrangian.

EP can be seen to be based on variational framework, using Bethe-like entropy and convex outer bound for the mean parameters.

When base distribution is unaries and $\Phi^i$ is the edges of a graph, we in fact get standard Bethe approximation, and standard sum-product LBP.

Moment matching of EP can be seen as striving for solution of associated Lagrangian.

Lost of flexibility here, depending on what the base distribution is (e.g., could be a $k$-tree or any other structure).

Can also be done for Gaussian mixture models.
EP as variational: Summary of key points

- Fixed points of EP exist assuming Lagrangian form has at least one optimum.
- No guarantees that EP will converge, but if it does it will be at a stationary point of the Lagrangian.
- EP can be seen to be based on variational framework, using Bethe-like entropy and convex outer bound for the mean parameters.
- When base distribution is unaries and $\Phi^i$ is the edges of a graph, we in fact get standard Bethe approximation, and standard sum-product LBP.
- Moment matching of EP can be seen as striving for solution of associated Lagrangian.
- Lost of flexibility here, depending on what the base distribution is (e.g., could be a $k$-tree or any other structure).
- Can also be done for Gaussian mixture models.
- Many more details, variations, and possible roads to new research. See text and also see Tom Minka’s papers.

So far, we have been using an outer bound on $\mathcal{M}$. 
So far, we have been using an outer bound on $\mathcal{M}$.

In mean-field methods, we use an “inner bound”, a subset of $\mathcal{M}$ constructed so as to make the optimization of $A(\theta)$ easier.
Mean Field

- So far, we have been using an outer bound on \( \mathcal{M} \).
- In mean-field methods, we use an “inner bound”, a subset of \( \mathcal{M} \) constructed so as to make the optimization of \( A(\theta) \) easier.
- Since subset, we get immediate bound on \( A(\theta) \).
So far, we have been using an outer bound on $\mathcal{M}$.

In mean-field methods, we use an “inner bound”, a subset of $\mathcal{M}$ constructed so as to make the optimization of $A(\theta)$ easier.

Since subset, we get immediate bound on $A(\theta)$.

Key: we based the inner bound on a “tractable family” like a 1-tree or even a 0-tree (all independent) so that the variational problem can be computed efficiently.
So far, we have been using an outer bound on $\mathcal{M}$.

In mean-field methods, we use an “inner bound”, a subset of $\mathcal{M}$ constructed so as to make the optimization of $A(\theta)$ easier.

Since subset, we get immediate bound on $A(\theta)$.

Key: we based the inner bound on a “tractable family” like a 1-tree or even a 0-tree (all independent) so that the variational problem can be computed efficiently.

Convexity is often lost still, however.
Tractable Families

- We have graph $G = (V, E)$ which is intractable and we find a spanning subgraph (recall, spanning = all nodes, subgraph = subset of edges), i.e., $F = (V, E_F)$ where $E_F \subseteq E$. 
Tractable Families

- We have graph $G = (V, E)$ which is intractable and we find a spanning subgraph (recall, spanning = all nodes, subgraph = subset of edges), i.e., $F = (V, E_F)$ where $E_F \subseteq E$.

- Simplest example: $F = (V, \emptyset)$.
Tractable Families

- We have graph $G = (V, E)$ which is intractable and we find a **spanning subgraph** (recall, spanning = all nodes, subgraph = subset of edges), i.e., $F = (V, E_F)$ where $E_F \subseteq E$.
- Simplest example: $F = (V, \emptyset)$ all independence model.
We have graph $G = (V, E)$ which is intractable and we find a spanning subgraph (recall, spanning = all nodes, subgraph = subset of edges), i.e., $F = (V, E_F)$ where $E_F \subseteq E$.

- Simplest example: $F = (V, \emptyset)$ all independence model.
- Tree example: $F = (V, E_T)$ where edges $E_T \subset E$ constitute a spanning tree.
Tractable Families

- We have graph $G = (V, E)$ which is intractable and we find a spanning subgraph (recall, spanning = all nodes, subgraph = subset of edges), i.e., $F = (V, E_F)$ where $E_F \subseteq E$.

- Simplest example: $F = (V, \emptyset)$ all independence model.

- Tree example: $F = (V, E_T)$ where edges $E_T \subset E$ constitute a spanning tree.

- Exponential family, sufficient statistics $\phi = (\phi_\alpha, \alpha \in \mathcal{I})$ associated with this family $\mathcal{I}(F) \subseteq \mathcal{I}$. These are the statistics that need respect the Markov properties of only the subgraph $F$. 
Tractable Families

- We have graph $G = (V, E)$ which is intractable and we find a spanning subgraph (recall, spanning = all nodes, subgraph = subset of edges), i.e., $F = (V, E_F)$ where $E_F \subseteq E$.
- Simplest example: $F = (V, \emptyset)$ all independence model.
- Tree example: $F = (V, E_T)$ where edges $E_T \subset E$ constitute a spanning tree.
- Exponential family, sufficient statistics $\phi = (\phi_\alpha, \alpha \in \mathcal{I})$ associated with this family $\mathcal{I}(F) \subseteq \mathcal{I}$. These are the statistics that need respect the Markov properties of only the subgraph $F$.
- $\Omega$ gets smaller too. The parameters that respect $F$ are of the form:

$$\mathbb{R}^{\mathcal{I}} \ni \Omega(F) \triangleq \{ \theta \in \Omega | \theta_\alpha = 0 \ \forall \alpha \in \mathcal{I} \setminus \mathcal{I}(F) \} \subseteq \Omega \quad (16.36)$$
Tractable Families

- We have graph $G = (V, E)$ which is intractable and we find a spanning subgraph (recall, spanning = all nodes, subgraph = subset of edges), i.e., $F = (V, E_F)$ where $E_F \subseteq E$.

- Simplest example: $F = (V, \emptyset)$ all independence model.

- Tree example: $F = (V, E_T)$ where edges $E_T \subset E$ constitute a spanning tree.

- Exponential family, sufficient statistics $\phi = (\phi_\alpha, \alpha \in \mathcal{I})$ associated with this family $\mathcal{I}(F) \subseteq \mathcal{I}$. These are the statistics that need respect the Markov properties of only the subgraph $F$.

- $\Omega$ gets smaller too. The parameters that respect $F$ are of the form:

  $$\mathbb{R}^{\lvert \mathcal{I} \rvert} \ni \Omega(F') \triangleq \{ \theta \in \Omega | \theta_\alpha = 0 \ \forall \alpha \in \mathcal{I} \setminus \mathcal{I}(F) \} \subseteq \Omega$$ (16.36)

  notice, all parameters associated with sufficient statistic not in $\mathcal{I}(F)$ are set to zero, those statistics are nonexistent in $F$. 
Tractable Families

- We have graph $G = (V, E)$ which is intractable and we find a **spanning subgraph** (recall, spanning = all nodes, subgraph = subset of edges), i.e., $F = (V, E_F)$ where $E_F \subseteq E$.

- Simplest example: $F = (V, \emptyset)$ all independence model.

- Tree example: $F = (V, E_T)$ where edges $E_T \subset E$ constitute a spanning tree.

- Exponential family, sufficient statistics $\phi = (\phi_\alpha, \alpha \in I)$ associated with this family $I(F) \subseteq I$. These are the statistics that need respect the Markov properties of only the subgraph $F$.

- $\Omega$ gets smaller too. The parameters that respect $F$ are of the form:

  $$\mathbb{R}^{|I|} \ni \Omega(F) \triangleq \{\theta \in \Omega | \theta_\alpha = 0 \ \forall \alpha \in I \setminus I(F)\} \subseteq \Omega$$

  (16.36)

  notice, all parameters associated with sufficient statistic not in $I(F)$ are set to zero, those statistics are nonexistent in $F$.

- If parameter was not zero, model would not respect the family of $F$. 
Tractable Subgraphs: All Independent Example

- Ex: MRF with potential functions for nodes and edges.
Tractable Subgraphs: All Independent Example

- **Ex:** MRF with potential functions for nodes and edges.
- **For each** $(s, t) \in E(G)$, **we have** \( \theta_{(s,t)} \).
Tractable Subgraphs: All Independent Example

- Ex: MRF with potential functions for nodes and edges.
- For each \((s, t) \in E(G)\), we have \(\theta_{(s,t)}\).
- \(F_0 = (V, \emptyset)\) which yields

\[
\Omega(F_0) = \{ \theta \in \Omega | \theta_{(s,t)} = 0 \ \forall (s, t) \in E(G) \} \tag{16.37}
\]
Ex: MRF with potential functions for nodes and edges.

For each \((s, t) \in E(G)\), we have \(\theta_{(s,t)}\).

\(F_0 = (V, \emptyset)\) which yields

\[
\Omega(F_0) = \{ \theta \in \Omega | \theta_{(s,t)} = 0 \ \forall (s, t) \in E(G) \} \quad (16.37)
\]

This is the all independence model, giving family of distributions

\[
p_\theta(x) = \prod_{s \in V} p(x_s ; \theta_s) \quad (16.38)
\]
Tractable Subgraphs: Tree Example

- Ex: MRF with potential functions for nodes and edges.
Tractable Subgraphs: Tree Example

- **Ex:** MRF with potential functions for nodes and edges.
- For each $(s, t) \in E(G)$, we have $\theta_{(s,t)}$. 

\[
\Omega(F) = \{ \theta \in \Omega | \theta(s,t) = 0 \quad \forall (s,t) \notin T \} \tag{16.39}
\]

This gives a tree-dependent family
\[
p_{\theta}(x) = \prod_{s \in V} p(x_s; \theta_s) \prod_{(s,t) \in T} p(x_s, x_t; \theta_{st}) p(x_s; \theta_s) p(x_t; \theta_t) \tag{16.40}
\]
Tractable Subgraphs: Tree Example

- Ex: MRF with potential functions for nodes and edges.
- For each \((s, t) \in E(G)\), we have \(\theta_{(s,t)}\).
- \(F_T = (V, T)\) where \(T \subset E\) are edges that constitute a spanning tree of \(G\), giving

\[
\Omega(F_0) = \{\theta \in \Omega | \theta_{(s,t)} = 0 \ \forall (s, t) \notin T\} \quad (16.39)
\]
Tractable Subgraphs: Tree Example

- **Ex:** MRF with potential functions for nodes and edges.
- For each \((s, t) \in E(G)\), we have \(\theta_{(s,t)}\).
- \(F_T = (V, T)\) where \(T \subset E\) are edges that constitute a spanning tree of \(G\), giving

\[
\Omega(F_0) = \{ \theta \in \Omega | \theta_{(s,t)} = 0 \ \forall (s, t) \notin T \} \tag{16.39}
\]

- This gives a tree-dependent family

\[
p_\theta(x) = \prod_{s \in V} p(x_s; \theta_s) \prod_{(s,t) \in T} \frac{p(x_s, x_t; \theta_{st})}{p(x_s; \theta_s)p(x_t; \theta_t)} \tag{16.40}
\]
Before, we had $\mathcal{M}(G; \phi)(= \mathcal{M}_G(G; \phi))$, all possible mean parameters associated with $G$ and associated set of sufficient statistics $\phi$. 
Before, we had $\mathcal{M}(G; \phi)(= \mathcal{M}_G(G; \phi))$, all possible mean parameters associated with $G$ and associated set of sufficient statistics $\phi$.

For a given subgraph $F$, we only consider those mean parameters possible under such models. I.e.,

$$\mathcal{M}_F(G; \phi) = \left\{ \mu \in \mathbb{R}^d | \mu = \mathbb{E}_\theta[\phi(x)] \text{ for some } \theta \in \Omega(F) \right\}$$  \hspace{1cm} (16.41)
Before, we had $\mathcal{M}(G; \phi)(= \mathcal{M}_G(G; \phi))$, all possible mean parameters associated with $G$ and associated set of sufficient statistics $\phi$.

For a given subgraph $F$, we only consider those mean parameters possible under such models. I.e.,

$$\mathcal{M}_F(G; \phi) = \left\{ \mu \in \mathbb{R}^d | \mu = \mathbb{E}_\theta[\phi(x)] \text{ for some } \theta \in \Omega(F) \right\} \quad (16.41)$$

Therefore, since $\theta \in \Omega(F) \subseteq \Omega$, we have that

$$\mathcal{M}_F^\circ(G; \phi) \subseteq \mathcal{M}^\circ(G; \phi) \quad (16.42)$$

and so $\mathcal{M}_F^\circ(G; \phi)$ is an inner approximation of the set of realizable mean parameters.
Before, we had \( \mathcal{M}(G; \phi) (= \mathcal{M}_G(G; \phi)) \), all possible mean parameters associated with \( G \) and associated set of sufficient statistics \( \phi \).

For a given subgraph \( F \), we only consider those mean parameters possible under such models. I.e.,

\[
\mathcal{M}_F(G; \phi) = \left\{ \mu \in \mathbb{R}^d \mid \mu = \mathbb{E}_{\theta}[\phi(x)] \text{ for some } \theta \in \Omega(F) \right\} \tag{16.41}
\]

Therefore, since \( \theta \in \Omega(F) \subseteq \Omega \), we have that

\[
\mathcal{M}^\circ_F(G; \phi) \subseteq \mathcal{M}^\circ(G; \phi) \tag{16.42}
\]

and so \( \mathcal{M}^\circ_F(G; \phi) \) is an inner approximation of the set of realizable mean parameters.

Shorthand notation: \( M^\circ_F(G) = M^\circ_F(G; \phi) \) and \( M^\circ(G) = M^\circ(G; \phi) \).
Mean field variational lower bound

- Mean field methods generate lower bounds on their estimated $A(\theta)$ and approximate mean parameters $\mu = \mathbb{E}_\theta[\phi(X)]$.

Proposition 16.5.1 (mean field lower bound)

Any mean parameter $\mu \in \mathcal{M}^\circ$ yields a lower bound on the cumulant function:

$$ A(\theta) \geq \langle \theta, \mu \rangle - A^*(\mu) \quad (16.43) $$

Moreover, equality holds if and only if $\theta$ and $\mu$ are dually coupled (i.e., $\mu = \mathbb{E}_\theta[\phi(X)]$).
Mean field variational lower bound

Proof.

1. On the one hand, obvious due to $A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\}$
Mean field variational lower bound

Proof.

- On the one hand, obvious due to $A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\}$

- More traditional proof, let $q$ be any distribution that satisfies moment matching $E_q[\phi(X)] = \mu$, then:

$$A(\theta) = \log \int_{\mathcal{X}^m} q(x) \frac{\exp \langle \theta, \phi(x) \rangle}{q(x)} \nu(dx)$$ \hspace{1cm} (16.44)

$$\geq \int_{\mathcal{X}^m} q(x) [\langle \theta, \phi(x) \rangle - \log q(x)] \nu(dx)$$ \hspace{1cm} (16.45)

$$= \langle \theta, E_q[\phi(X)] \rangle - H(q) = \langle \theta, \mu \rangle - H(q)$$ \hspace{1cm} (16.46)
Mean field variational lower bound

Proof.

- On the one hand, obvious due to $A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\}$

- More traditional proof, let $q$ be any distribution that satisfies moment matching $\mathbb{E}_q[\phi(X)] = \mu$, then:

  \[
  A(\theta) = \log \int_{\mathcal{X}^m} q(x) \frac{\exp \langle \theta, \phi(x) \rangle}{q(x)} \nu(dx) \]

  \[
  \geq \int_{\mathcal{X}^m} q(x)[\langle \theta, \phi(x) \rangle - \log q(x)] \nu(dx) \]

  \[
  = \langle \theta, E_q[\phi(X)] \rangle - H(q) = \langle \theta, \mu \rangle - H(q)
  \]

- If we optimize $q$ over all $\mathcal{M}(G)$, then we’ll get equality.
**Mean field variational lower bound**

**Proof.**

- On the one hand, obvious due to $A(\theta) = \sup_{\mu \in \mathcal{M}} \{ \langle \theta, \mu \rangle - A^*(\mu) \}$

- More traditional proof, let $q$ be any distribution that satisfies moment matching $\mathbb{E}_q[\phi(X)] = \mu$, then:

\[
A(\theta) = \log \int_{\mathcal{X}_m^}\ q(x) \frac{\exp \langle \theta, \phi(x) \rangle}{q(x)} \nu(dx) \tag{16.44}
\]

\[
\geq \int_{\mathcal{X}_m^}\ q(x) [\langle \theta, \phi(x) \rangle - \log q(x)] \nu(dx) \tag{16.45}
\]

\[
= \langle \theta, E_q[\phi(X)] \rangle - H(q) = \langle \theta, \mu \rangle - H(q) \tag{16.46}
\]

- If we optimize $q$ over all $\mathcal{M}(G)$, then we’ll get equality.

- If we optimize $q$ over a subset of $\mathcal{M}(G)$ (e.g., such as $\mathcal{M}_F(G)$), then we’ll get inequality.
Tractable Dual

- Normally dual $A^*(\mu) = \sup_{\theta \in \Omega} (\langle \theta, \mu \rangle - A(\theta))$ is intractable or unavailable, but key idea is that if $\mu \in \mathcal{M}_F(G)$ it will be possible to compute easily.
**Tractable Dual**

- Normally dual $A^*(\mu) = \sup_{\theta \in \Omega} (\langle \theta, \mu \rangle - A(\theta))$ is intractable or unavailable, but key idea is that if $\mu \in \mathcal{M}_F(G)$ it will be possible to compute easily.

- Thus, goal of mean field (from variational approximation perspective) is to form $A_{MF}(\theta)$ where:

\[
A(\theta) \geq \max_{\mu \in \mathcal{M}_F(G)} \{ \langle \mu, \theta \rangle - A^*_F(\mu) \} \triangleq A_{MF}(\theta) \quad (16.47)
\]

where $A^*_F(\mu)$ corresponds to dual function restricted to inner bound set $\mathcal{F}(G)$. I.e., when we expand $A^*_F(\mu)$, we can take advantage of the fact that $\mu$ is restricted in all cases, so $A^*_F(\mu)$ might be greatly simplified relative to $A^*(\mu)$. 

Tractable Dual

- Normally dual $A^*(\mu) = \sup_{\theta \in \Omega} (\langle \theta, \mu \rangle - A(\theta))$ is intractable or unavailable, but key idea is that if $\mu \in \mathcal{M}_F(G)$ it will be possible to compute easily.

- Thus, goal of mean field (from variational approximation perspective) is to form $A_{MF}(\theta)$ where:

\[
A(\theta) \geq \max_{\mu \in \mathcal{M}_F(G)} \{ \langle \mu, \theta \rangle - A^*_F(\mu) \} \triangleq A_{MF}(\theta) \tag{16.47}
\]

where $A^*_F(\mu)$ corresponds to dual function restricted to inner bound set $\mathcal{F}(G)$. I.e., when we expand $A^*_F(\mu)$, we can take advantage of the fact that $\mu$ is restricted in all cases, so $A^*_F(\mu)$ might be greatly simplified relative to $A^*(\mu)$.

- Note, for $\mu \in \mathcal{M}_F(G)$, $A^*_F(\mu)$ is not an approximation, rather it is just easy to compute.
Given two distributions \( p, q \), KL-Divergence of \( p \) w.r.t. \( q \) is defined as

\[
D(q \| p) = \int_{\mathcal{X}} q(x) \left[ \log \frac{q(x)}{p(x)} \right] \nu(dx)
\]

(16.48)
Given two distributions \( p, q \), KL-Divergence of \( p \) w.r.t. \( q \) is defined as

\[
D(q\|p) = \int_{\mathcal{X}^m} q(x) \left[ \log \frac{q(x)}{p(x)} \right] \nu(dx)
\]  

(16.48)

In summation form, we have

\[
D(q\|p) = \sum_{x \in \mathcal{X}^m} q(x) \left[ \log \frac{q(x)}{p(x)} \right]
\]  

(16.49)
Given two distributions $p$, $q$, KL-Divergence of $p$ w.r.t. $q$ is defined as

$$D(q\|p) = \int_{x^m} q(x) \left[ \log \frac{q(x)}{p(x)} \right] \nu(dx) \quad (16.48)$$

In summation form, we have

$$D(q\|p) = \sum_{x \in x^m} q(x) \left[ \log \frac{q(x)}{p(x)} \right] \quad (16.49)$$

For exponential models this takes on some interesting forms, and moreover, we can see the variational approximation above as a KL-divergence minimization problem.
Given two distributions $p, q$, KL-Divergence of $p$ w.r.t. $q$ is defined as

$$D(q||p) = \int_{\mathcal{X}^m} q(x) \left[ \log \frac{q(x)}{p(x)} \right] \nu(dx)$$  \hspace{1cm} (16.48)

In summation form, we have

$$D(q||p) = \sum_{x \in \mathcal{X}^m} q(x) \left[ \log \frac{q(x)}{p(x)} \right]$$  \hspace{1cm} (16.49)

For exponential models this takes on some interesting forms, and more over, we can see the variational approximation above as a KL-divergence minimization problem.

Recall, exponential models can be parameterized using canonical parameters $\theta$ or mean parameters $\mu$. We will use notational shortcuts: $D(\theta^1||\theta^2) \equiv D(p_{\theta^1}||p_{\theta^2})$, $D(\mu^1||\mu^2) \equiv D(p_{\mu^1}||p_{\mu^2})$, and even $D(\mu^1||\theta^2) \equiv D(p_{\mu^1}||p_{\theta^2})$. 
Mean field, KL-Divergence, Exponential Model Families

- Consider $\theta^1, \theta^2 \in \Omega$
Consider $\theta^1, \theta^2 \in \Omega$

Let $D(\theta^1 || \theta^2)$ have aforementioned meaning (KL-divergence between the two corresponding distributions), and let $\mu^i = \mathbb{E}_{\theta^i}[\phi(X)]$, 

$$\text{(16.50)}$$

$$\text{(16.51)}$$

$$\text{(16.52)}$$
Consider $\theta^1, \theta^2 \in \Omega$

Let $D(\theta^1 \| \theta^2)$ have aforementioned meaning (KL-divergence between the two corresponding distributions), and let $\mu^i = \mathbb{E}_{\theta^i}[\phi(X)]$

Then we have a Bregman divergence form:

$$D(\theta^1 \| \theta^2) = \mathbb{E}_{\theta^1} \left[ \log \frac{p_{\theta^1}(x)}{p_{\theta^2}(x)} \right]$$

$$= A(\theta^2) - A(\theta^1) - \langle \mu^1, \theta^2 - \theta^1 \rangle$$

$$= A(\theta^2) - \left[ A(\theta^1) + \langle \nabla A(\theta^1), \theta^2 - \theta^1 \rangle \right]$$

$$D(\theta^1 \| \theta^2) = A(\theta^2) - A(\theta^1) - \langle \mu^1, \theta^2 - \theta^1 \rangle$$

$$= A(\theta^2) - \left[ A(\theta^1) + \langle \nabla A(\theta^1), \theta^2 - \theta^1 \rangle \right]$$
Purely dual form of KL divergence can be formed as well, i.e.,

\[ D(\theta^1||\theta^2) = D(\mu^1||\mu^2) = A^*(\mu^1) - A^*(\mu^2) - \langle \theta^2, \mu^1 - \mu^2 \rangle \] (16.53)
Purely dual form of KL divergence can be formed as well, i.e.,

\[ D(\theta^1 || \theta^2) = D(\mu^1 || \mu^2) = A^*(\mu^1) - A^*(\mu^2) - \langle \theta^2, \mu^1 - \mu^2 \rangle \]  

Dual Bregman form
Mean field, KL-Divergence, Exponential Model Families

- Mixed/hybrid form of KL in terms of dual
Mixed/hybrid form of KL in terms of dual

We can also write the KL as:

\[ D(\theta^1||\theta^2) = D(\mu^1||\theta^2) = A(\theta^2) + A^*(\mu^1) - \langle \mu^1, \theta^2 \rangle \]  \hspace{1cm} (16.54)

which comes from dual expression \( A^*(\mu^1) = \langle \theta^1, \mu^1 \rangle - A(\theta^1) \) for dually coupled parameters \( \mu^1 = E_{\theta^1}[\phi(X)] \).
Mixed/hybrid form of KL in terms of dual

We can also write the KL as:

\[ D(\theta^1 || \theta^2) = D(\mu^1 || \theta^2) = A(\theta^2) + A^*(\mu^1) - \langle \mu^1, \theta^2 \rangle \quad (16.54) \]

which comes from dual expression \( A^*(\mu^1) = \langle \theta^1, \mu^1 \rangle - A(\theta^1) \) for dually coupled parameters \( \mu^1 = \mathbb{E}_{\theta^1}[\phi(X)] \).

In particular, this equation (variational expression for the cumulant):

\[ A(\theta) = \sup_{\mu \in \mathcal{M}} \{ \langle \theta, \mu \rangle - A^*(\mu) \} \quad (??) \]
Mean field, KL-Divergence, Exponential Model Families

- Mixed/hybrid form of KL in terms of dual
  - We can also write the KL as:
    \[
    D(\theta^1||\theta^2) = D(\mu^1||\theta^2) = A(\theta^2) + A^*(\mu^1) - \langle \mu^1, \theta^2 \rangle
    \] (16.54)

  which comes from dual expression
  \[
  A^*(\mu^1) = \langle \theta^1, \mu^1 \rangle - A(\theta^1)
  \]
  for dually coupled parameters
  \[
  \mu^1 = \mathbb{E}_{\theta^1}[\phi(X)].
  \]

- In particular, this equation (variational expression for the cumulant):
  \[
  A(\theta) = \sup_{\mu \in \mathcal{M}} \{ \langle \theta, \mu \rangle - A^*(\mu) \}
  \] (??)

  ...can be written as:
  \[
  \inf_{\mu \in \mathcal{M}} \{ A(\theta) + A^*(\mu) - \langle \theta, \mu \rangle \} = \inf_{\mu \in \mathcal{M}} D(\mu||\theta) = 0 \] (16.55)
Since

$$\inf_{\mu \in \mathcal{M}} \{ A(\theta) + A^*(\mu) - \langle \theta, \mu \rangle \} = \inf_{\mu \in \mathcal{M}} D(\mu||\theta) = 0$$  \hspace{1cm} (16.55)
Since

\[
\inf_{\mu \in \mathcal{M}} \left\{ A(\theta) + A^*(\mu) - \langle \theta, \mu \rangle \right\} = \inf_{\mu \in \mathcal{M}} D(\mu||\theta) = 0 \quad (16.55)
\]

Thus, solving the mean-field variational problem of:

\[
\max_{\mu \in \mathcal{M}_F(G)} \left\{ \langle \mu, \theta \rangle - A^*_F(\mu) \right\} \quad (16.47)
\]

is identical to minimizing KL Divergence \( D(\mu||\theta) \) subject to constraint \( \mu \in \mathcal{M}_F(G) \).
Since

\[
\inf_{\mu \in \mathcal{M}} \{ A(\theta) + A^*(\mu) - \langle \theta, \mu \rangle \} = \inf_{\mu \in \mathcal{M}} D(\mu || \theta) = 0 \quad (16.55)
\]

Thus, solving the mean-field variational problem of:

\[
\max_{\mu \in \mathcal{M}_F(G)} \{ \langle \mu, \theta \rangle - A^*_F(\mu) \} \quad (16.47)
\]

is identical to minimizing KL Divergence \( D(\mu || \theta) \) subject to constraint \( \mu \in \mathcal{M}_F(G) \).

I.e., mean field can be seen as finding the best approximation, in terms of this particular KL-divergence, to \( p_\theta \), over a family of “nice” distributions \( M_F(G) \).
A classic example of mean-field (goes back to statistical physics)
A classic example of mean-field (goes back to statistical physics)

Mean parameters for Ising: \( \mu_s = \mathbb{E}[X_s] = p(X_s = 1), \)
\( \mu_{st} = \mathbb{E}[X_sX_t] = p(X_s = 1, X_t = 1), \) thus \( \mu \in \mathbb{R}^{V+|E|}. \)
A classic example of mean-field (goes back to statistical physics)

Mean parameters for Ising: \( \mu_s = \mathbb{E}[X_s] = p(X_s = 1) \), \( \mu_{st} = \mathbb{E}[X_sX_t] = p(X_s = 1, X_t = 1) \), thus \( \mu \in \mathbb{R}^{|V|+|E|} \).

Let \( F_0 = (V, \emptyset) \) be our mean field approximation family. Thus,

\[
\mathcal{M}_{F_0}(G) = \left\{ \mu \in \mathbb{R}^{|V|+|E|} \mid 0 \leq \mu_s \leq 1 \ \forall s \in V, \ \text{and} \ \mu_{st} = \mu_s \mu_t \ \forall \right\}
\]
Naïve Mean field for Ising Model

- A classic example of mean-field (goes back to statistical physics)
- Mean parameters for Ising: $\mu_s = \mathbb{E}[X_s] = p(X_s = 1)$,
  $\mu_{st} = \mathbb{E}[X_sX_t] = p(X_s = 1, X_t = 1)$, thus $\mu \in \mathbb{R}^{|V|+|E|}$.
- Let $F_0 = (V, \emptyset)$ be our mean field approximation family. Thus,

$$\mathcal{M}_{F_0}(G) = \left\{ \mu \in \mathbb{R}^{|V|+|E|} \mid 0 \leq \mu_s \leq 1 \ \forall s \in V, \text{ and } \mu_{st} = \mu_s \mu_t \ \forall \right\}$$

- Key is that for $\mu \in \mathcal{M}_{F_0}(G)$, dual is not hard to calculate, that is

$$-A^*_F(\mu) = \sum_{s \in V} H_s(\mu_s) \quad (16.56)$$

which are sum of unary entropy terms, very cheap.
Naïve Mean field for Ising Model

- A classic example of mean-field (goes back to statistical physics)
- Mean parameters for Ising: $\mu_s = \mathbb{E}[X_s] = p(X_s = 1)$, $\mu_{st} = \mathbb{E}[X_sX_t] = p(X_s = 1, X_t = 1)$, thus $\mu \in \mathbb{R}^{|V|+|E|}$.
- Let $F_0 = (V, \emptyset)$ be our mean field approximation family. Thus,

$$
\mathcal{M}_{F_0}(G) = \left\{ \mu \in \mathbb{R}^{|V|+|E|} \mid 0 \leq \mu_s \leq 1 \ \forall s \in V, \text{ and } \mu_{st} = \mu_s \mu_t \ \forall \right\}
$$

- Key is that for $\mu \in \mathcal{M}_{F_0}(G)$, dual is not hard to calculate, that is

$$
-A^*_F(\mu) = \sum_{s \in V} H_s(\mu_s) \quad (16.56)
$$

which are sum of unary entropy terms, very cheap.
- Moreover, polytope for $M_{F_0}(G)$ is also very simple, namely the hypercube $[0, 1]^m$.
We get variational lower bound problem

\[
A(\theta) \geq \max_{(\mu_1,...,\mu_m) \in [0,1]^m} \left\{ \sum_{s \in V} \theta_s \mu_s + \sum_{(s,t) \in E} \theta_{st} \mu_s \mu_t + \sum_{s \in V} H_s(\mu_s) \right\}
\] (16.57)
Naive Mean field for Ising Model

- We get variational lower bound problem

\[
A(\theta) \geq \max_{(\mu_1, ..., \mu_m) \in [0,1]^m} \left\{ \sum_{s \in V} \theta_s \mu_s + \sum_{(s,t) \in E} \theta_{st} \mu_s \mu_t + \sum_{s \in V} H_s(\mu_s) \right\}
\] (16.57)

- Have constrained form of edge mean parameters \( \mu_{st} = \mu_s \mu_t \)
Naive Mean field for Ising Model

- We get variational lower bound problem

\[ A(\theta) \geq \max_{(\mu_1,\ldots,\mu_m) \in [0,1]^m} \left\{ \sum_{s \in V} \theta_s \mu_s + \sum_{(s,t) \in E} \theta_{st} \mu_s \mu_t + \sum_{s \in V} H_s(\mu_s) \right\} \]  

(16.57)

- Have constrained form of edge mean parameters \( \mu_{st} = \mu_s \mu_t \)
- \((\mu_1,\ldots,\mu_m) \in [0,1]^m\) is \(m\)-D hypercube.
Naive Mean field for Ising Model

- We get variational lower bound problem

\[
A(\theta) \geq \max_{(\mu_1, \ldots, \mu_m) \in [0,1]^m} \left\{ \sum_{s \in V} \theta_s \mu_s + \sum_{(s,t) \in E} \theta_{st} \mu_s \mu_t + \sum_{s \in V} H_s(\mu_s) \right\}
\]  

(16.57)

- Have constrained form of edge mean parameters \( \mu_{st} = \mu_s \mu_t \)
- \((\mu_1, \ldots, \mu_m) \in [0,1]^m\) is \(m\)-D hypercube.
- Once again, we have a non-convex problem.
We get variational lower bound problem

\[
A(\theta) \geq \max_{(\mu_1, \ldots, \mu_m) \in [0, 1]^m} \left\{ \sum_{s \in V} \theta_s \mu_s + \sum_{(s, t) \in E} \theta_{st} \mu_s \mu_t + \sum_{s \in V} H_s(\mu_s) \right\}
\]

Have constrained form of edge mean parameters \( \mu_{st} = \mu_s \mu_t \)

\((\mu_1, \ldots, \mu_m) \in [0, 1]^m \) is \(m\)-D hypercube.

Once again, we have a non-convex problem.

One way to optimize is to do coordinate ascent (given otherwise fixed vector, optimize one value at a time).
Naive Mean field for Ising Model

- We get variational lower bound problem

\[
A(\theta) \geq \max_{(\mu_1, \ldots, \mu_m) \in [0,1]^m} \left\{ \sum_{s \in V} \theta_s \mu_s + \sum_{(s,t) \in E} \theta_{st} \mu_s \mu_t + \sum_{s \in V} H_s(\mu_s) \right\}
\]

(16.57)

- Have constrained form of edge mean parameters \( \mu_{st} = \mu_s \mu_t \)
- \((\mu_1, \ldots, \mu_m) \in [0,1]^m\) is \(m\)-D hypercube.
- Once again, we have a non-convex problem.
- One way to optimize is to do coordinate ascent (given otherwise fixed vector, optimize one value at a time).
- If each coordinate optimization is optimal, we’ll get a stationary point.
Naive Mean field for Ising Model

- coordinate ascent: choose some $s$ and optimize $\mu_s$ fixing all $\mu_t$ for $t \neq s$. 

Taking derivatives w.r.t. $\mu_s$, we get the following update rule for element $\mu_s$

$$
\mu_s \leftarrow \sigma \left( \theta_s + \sum_{t \in N(s)} \theta_{st} \mu_t \right)
$$

where

$$
\sigma(z) = \frac{1}{1 + \exp(-z)}
$$

is the sigmoid (logistic) function.

This is the standard mean-field update that is quite well known, but derived from coordinate ascent optimization of a variational perspective of the problem. The variational approach indeed seems quite general and powerful.
Naive Mean field for Ising Model

- coordinate ascent: choose some $s$ and optimize $\mu_s$ fixing all $\mu_t$ for $t \neq s$.

- Taking derivatives w.r.t. $\mu_s$, we get the following update rule for element $\mu_s$

\[
\mu_s \leftarrow \sigma \left( \theta_s + \sum_{t \in N(s)} \theta_{st} \mu_t \right) \quad (16.58)
\]

where $\sigma(z) = [1 + \exp(-z)]^{-1}$ is the sigmoid (logistic) function.
Naive Mean field for Ising Model

- coordinate ascent: choose some $s$ and optimize $\mu_s$ fixing all $\mu_t$ for $t \neq s$.
- Taking derivatives w.r.t. $\mu_s$, we get the following update rule for element $\mu_s$

$$\mu_s \leftarrow \sigma \left( \theta_s + \sum_{t \in N(s)} \theta_{st} \mu_t \right)$$  \hspace{1cm} (16.58)

where $\sigma(z) = \left[ 1 + \exp(-z) \right]^{-1}$ is the sigmoid (logistic) function.

- This is the standard mean-field update that is quite well known, but derived from coordinate ascent optimization of a variational perspective of the problem.
Naive Mean field for Ising Model

- coordinate ascent: choose some \( s \) and optimize \( \mu_s \) fixing all \( \mu_t \) for \( t \neq s \).

- Taking derivatives w.r.t. \( \mu_s \), we get the following update rule for element \( \mu_s \)

\[
\mu_s \leftarrow \sigma \left( \theta_s + \sum_{t \in N(s)} \theta_{st} \mu_t \right) \tag{16.58}
\]

where \( \sigma(z) = \frac{1}{1 + \exp(-z)} \) is the sigmoid (logistic) function.

- This is the standard mean-field update that is quite well known, but derived from coordinate ascent optimization of a variational perspective of the problem.

- The variational approach indeed seems quite general and powerful.
Example of Lack of Convexity

- Consider simple two variable example \((X_1, X_2), X_i \in \{-1, +1\}\).
Example of Lack of Convexity

- Consider simple two variable example \((X_1, X_2), X_i \in \{-1, +1\}\).
- Exponential family form

\[
p_{\theta}(x) \propto \exp(\theta_1 x_1 + \theta_2 x_2 + \theta_{12} x_1 x_2) \quad (16.59)
\]

having mean parameters \(\mu_i = \mathbb{E}[X_i]\) and \(\mu_{12} = \mathbb{E}[X_1 X_2]\).
Example of Lack of Convexity

- Consider simple two variable example \((X_1, X_2), X_i \in \{-1, +1\}\).
- Exponential family form
  \[
p_\theta(x) \propto \exp(\theta_1 x_1 + \theta_2 x_2 + \theta_{12} x_1 x_2) \tag{16.59}\]
  having mean parameters \(\mu_i = \mathbb{E}[X_i]\) and \(\mu_{12} = \mathbb{E}[X_1 X_2]\).
- Impose constraint \(\mu_{12} = \mu_1 \mu_2\), we get mean field objective
  \[
f(\mu_1, \mu_2; \theta) = \theta_{12} \mu_1 \mu_2 + \theta_1 \mu_1 + \theta_2 \mu_2 + H(\mu_1) + H(\mu_2) \tag{16.60}\]
  where \(H(\mu) = -\frac{1}{2} (1 + \mu) \log \frac{1}{2} (1 + \mu) - \frac{1}{2} (1 - \mu) \log \frac{1}{2} (1 - \mu)\)

Note that \(p(X_i = +1) = \frac{1}{2} (1 + \mu_i)\)
Example of Lack of Convexity

- Consider simple two variable example \((X_1, X_2)\), \(X_i \in \{-1, +1\}\).
- Exponential family form

\[
p_\theta(x) \propto \exp(\theta_1 x_1 + \theta_2 x_2 + \theta_{12} x_1 x_2)
\]

having mean parameters \(\mu_i = \mathbb{E}[X_i]\) and \(\mu_{12} = \mathbb{E}[X_1 X_2]\).
- Impose constraint \(\mu_{12} = \mu_1 \mu_2\), we get mean field objective

\[
f(\mu_1, \mu_2; \theta) = \theta_{12} \mu_1 \mu_2 + \theta_1 \mu_1 + \theta_2 \mu_2 + H(\mu_1) + H(\mu_2)
\]

where \(H(\mu_i) = -\frac{1}{2} (1 + \mu_i) \log \frac{1}{2} (1 + \mu_i) - \frac{1}{2} (1 - \mu_i) \log \frac{1}{2} (1 - \mu_i)\)

- Consider sub-models of the form:

\[
(\theta_1, \theta_2, \theta_{12}) = \left(0, 0, \frac{1}{4} \log \frac{q}{1 - q} \right) \triangleq \theta(q)
\]

where \(q \in (0, 1)\) is a parameter such that, for any \(q\) we have \(\mathbb{E}[X_i] = 0\). It turns out that in this form, we have \(q = p(X_1 = X_2)\).
Example of Lack of Convexity

- Consider simple two variable example \((X_1, X_2), X_i \in \{-1, +1\}\).

- Exponential family form

\[
p_\theta(x) \propto \exp(\theta_1 x_1 + \theta_2 x_2 + \theta_{12} x_1 x_2) \tag{16.59}\]

having mean parameters \(\mu_i = \mathbb{E}[X_i]\) and \(\mu_{12} = \mathbb{E}[X_1 X_2]\).

- Impose constraint \(\mu_{12} = \mu_1 \mu_2\), we get mean field objective

\[
f(\mu_1, \mu_2; \theta) = \theta_{12} \mu_1 \mu_2 + \theta_1 \mu_1 + \theta_2 \mu_2 + H(\mu_1) + H(\mu_2) \tag{16.60}\]

where \(H(\mu_i) = -\frac{1}{2}(1 + \mu_i) \log \frac{1}{2}(1 + \mu_i) - \frac{1}{2}(1 - \mu_i) \log \frac{1}{2}(1 - \mu_i)\)

- Consider sub-models of the form:

\[
(\theta_1, \theta_2, \theta_{12}) = \left(0, 0, \frac{1}{4} \log \frac{q}{1 - q}\right) \triangleq \theta(q) \tag{16.61}\]

where \(q \in (0, 1)\) is a parameter such that, for any \(q\) we have \(\mathbb{E}[X_i] = 0\). It turns out that in this form, we have \(q = p(X_1 = X_2)\).

- Is mean field objective in this case convex for all \(q\)?
Lack of Convexity example

- For $q = 0.5$, objective $f(\mu_1, \mu_2; \theta(0.5))$ has global maximum at $(\mu_1, \mu_2) = (0, 0)$ so mean field is exact and convex. This corresponds to $p(X_1 = X_2) = 0$. When $q$ gets small, $f$ becomes non-convex, e.g., has multiple modes in figure.
Lack of Convexity example

- For $q = 0.5$, objective $f(\mu_1, \mu_2; \theta(0.5))$ has **global** maximum at $(\mu_1, \mu_2) = (0, 0)$ so mean field is exact and convex. This corresponds to $p(X_1 = X_2) = 0$.

- When $q$ gets small, $f$ becomes non-convex, e.g., has multiple modes in figure.
Lack of Convexity example

- For $q = 0.5$, objective $f(\mu_1, \mu_2; \theta(0.5))$ has **global** maximum at $(\mu_1, \mu_2) = (0, 0)$ so mean field is exact and convex. This corresponds to $p(X_1 = X_2) = 0$.
- When $q$ gets small, $f$ becomes non-convex, e.g., has multiple modes in figure.
Sources for Today’s Lecture