EE12A – Advanced Inference in Graphical Models
Fall 2011

Prof. Jeff Bilmes

University of Washington, Seattle
Department of Electrical Engineering
Fall Quarter, 2011
http://j.ee.washington.edu/~bilmes/classes/ee512a_fall_2011/

Lecture 15 - Nov 23th, 2011
Outstanding Reading

- Finish chapter 4 in the Wainwright and Jordan book.
- More project status reports (3rd status reports) are due tonight Wednesday evening (night before thanksgiving break) at 11:00pm via our dropbox (https://catalyst.uw.edu/collectit/dropbox/bilmes/17463). Remember, no more than one page!
We need to find one makeup lecture this term.

- **L1 (9/28):** Introduction, Families, Semantics
- **LX (9/30):** No class
- **L2 (10/5):** Trees, exact inference
- **L3 (10/7):** More on trees and inference.
- **L4 (10/12):** To tree or not to tree.
- **L5 (10/14):** All models lead to trees
- **L6 (10/19):** Decomposable, JT
- **L7 (10/21):** Inference on JTs
- **L8 (10/26):** JT Inference, semi-rings,
- **L9 (10/28):** time-space tradeoff, conditioning, LBP
- **L10 (11/2):** LBP, exp. f. models
- **L11 (11/4):** exp. f. models, marg poly
- **L12 (11/9):** pseudo marg, Bethe
- **LXX (11/11):** Veterans Day, no class
- **L13 (11/16):** Bethe, loop series
- **L14 (11/18):** loop series, EP
- **L15 (11/23):** EP
- **LXX (11/25):** Thanksgiving, no class
- **L16 (11/30):**
- **L17 (12/2):**
- **L18 (12/7):**
- **L19: (12/9):**
- **L20: ???:**
Finish chapter 4 from Wainwright & Jordan book.
Theorem 2.1 (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:

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

**(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 $\mu \in \mathcal{M}^\circ$ at moment matching conditions

$$
\mu = \int_{D_X} \phi(x)p_\theta(x)\nu(dx) = \mathbb{E}_\theta[\phi(X)] = \nabla A(\theta)
$$
Bethe Entropy Approximation

- We can perhaps just use this as an approximation, i.e., say that for any graph $G = (V, E)$ not nec. a tree.
Bethe Entropy Approximation

- We can perhaps just use this as an approximation, i.e., say that for any graph \( G = (V, E) \) not nec. a tree.

- That is, assuming that the distribution is structured over pairwise potential functions w.r.t. a graph \( G \), we can make an approximation to \(-A^*(\tau)\) based on equation that has same form, i.e.,

\[
-A^*(\tau) \approx H_{\text{Bethe}}(\tau) \triangleq \sum_{v \in V(G)} H_v(\tau_v) - \sum_{(s,t) \in E(G)} I_{st}(\tau_{st}) \quad (4)
\]
Bethe Entropy Approximation

- We can perhaps just use this as an approximation, i.e., say that for any graph $G = (V, E)$ not nec. a tree.

- That is, assuming that the distribution is structured over pairwise potential functions w.r.t. a graph $G$, we can make an approximation to $-A^*(\tau)$ based on equation that has same form, i.e.,

$$
-A^*(\tau) \approx H_{\text{Bethe}}(\tau) \triangleq \sum_{v \in V(G)} H_v(\tau_v) - \sum_{(s,t) \in E(G)} I_{st}(\tau_{st})
$$

- Key: $H_{\text{Bethe}}(\tau)$ is not necessarily concave as it is not a real entropy.
Bethe Entropy Approximation

- We can perhaps just use this as an approximation, i.e., say that for any graph \( G = (V, E) \) not nec. a tree.

- That is, assuming that the distribution is structured over pairwise potential functions w.r.t. a graph \( G \), we can make an approximation to \(-A^*(\tau)\) based on equation that has same form, i.e.,

\[
-A^*(\tau) \approx H_{\text{Bethe}}(\tau) \triangleq \sum_{v \in V(G)} H_v(\tau_v) - \sum_{(s,t) \in E(G)} I_{st}(\tau_{st})
\]  

(4)

- Key: \( H_{\text{Bethe}}(\tau) \) is not necessarily concave as it is not a real entropy.

- MI equation is not hard to compute \( O(r^2) \).

\[
I_{st}(\tau_{st}) = I_{st}(\tau_{st}(x_s, x_t))
\]

(5)

\[
= \sum_{x_s, x_t} \tau_{st}(x_s, x_t) \log \frac{\tau_{st}(x_s, x_t)}{\tau_s(x_s) \tau_t(x_t)}
\]

(6)
Bethe Entropy Approximation

- We can perhaps just use this as an approximation, i.e., say that for any graph $G = (V, E)$ not nec. a tree.

- That is, assuming that the distribution is structured over pairwise potential functions w.r.t. a graph $G$, we can make an approximation to $-A^*(\tau)$ based on equation that has same form, i.e.,

$$ -A^*(\tau) \approx H_{\text{Bethe}}(\tau) \triangleq \sum_{v \in V(G)} H_v(\tau_v) - \sum_{(s,t) \in E(G)} I_{st}(\tau_{st}) $$

- Key: $H_{\text{Bethe}}(\tau)$ is not necessarily concave as it is not a real entropy.

- MI equation is not hard to compute $O(r^2)$.

$$ I_{st}(\tau_{st}) = I_{st}(\tau_{st}(x_s, x_t)) $$

$$ = \sum_{x_s, x_t} \tau_{st}(x_s, x_t) \log \frac{\tau_{st}(x_s, x_t)}{\tau_s(x_s) \tau_t(x_t)} $$

- Lagrange multipliers corresponds to LBP messages, fixed points.
Bethe Variational Problem and LBP

Original variational representation of log partition function

\[ A(\theta) = \sup_{\mu \in \mathcal{M}} \{ \langle \theta, \mu \rangle - A^*(\mu) \} \]  \hspace{1cm} (7)
Bethe Variational Problem and LBP

Original variational representation of log partition function

\[ A(\theta) = \sup_{\mu \in \mathcal{M}} \{ \langle \theta, \mu \rangle - A^*(\mu) \} \]  (7)

Approximate variational representation of log partition function

\[ A_{\text{Bethe}}(\theta) = \sup_{\tau \in \mathcal{L}} \{ \langle \theta, \tau \rangle + H_{\text{Bethe}}(\tau) \} \]  (8)

\[ = \sup_{\tau \in \mathcal{L}} \left\{ \langle \theta, \tau \rangle + \sum_{v \in V(G)} H_v(\tau_v) - \sum_{(s,t) \in E(G)} I_{st}(\tau_{st}) \right\} \]  (9)
Bethe Variational Problem and LBP

Original variational representation of log partition function

\[ A(\theta) = \sup_{\mu \in \mathcal{M}} \{ \langle \theta, \mu \rangle - A^*(\mu) \} \]  \hspace{1cm} (7)

Approximate variational representation of log partition function

\[ A_{\text{Bethe}}(\theta) = \sup_{\tau \in \mathbb{L}} \{ \langle \theta, \tau \rangle + H_{\text{Bethe}}(\tau) \} \]  \hspace{1cm} (8)

\[ = \sup_{\tau \in \mathbb{L}} \left\{ \langle \theta, \tau \rangle + \sum_{v \in V(G)} H_v(\tau_v) - \sum_{(s,t) \in E(G)} I_{st}(\tau_{st}) \right\} \]  \hspace{1cm} (9)

- Exact when \( G = T \) but we do this for any \( G \), still commutable
Bethe Variational Problem and LBP

Original variational representation of log partition function

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

Approximate variational representation of log partition function

\[
A_{\text{Bethe}}(\theta) = \sup_{\tau \in \mathbb{I}} \{ \langle \theta, \tau \rangle + H_{\text{Bethe}}(\tau) \}
\]

\[
= \sup_{\tau \in \mathbb{I}} \left\{ \langle \theta, \tau \rangle + \sum_{v \in V(G)} H_v(\tau_v) - \sum_{(s,t) \in E(G)} I_{st}(\tau_{st}) \right\}
\]

- Exact when \( G = T \) but we do this for any \( G \), still commutable
- we get an approximate log partition function, and approximate (pseudo) marginals (in \( \mathbb{I} \)), but this is perhaps much easier to compute.
Bethe Variational Problem and LBP

Original variational representation of log partition function

\[ A(\theta) = \sup_{\mu \in \mathcal{M}} \{ \langle \theta, \mu \rangle - A^*(\mu) \} \]  (7)

Approximate variational representation of log partition function

\[ A_{\text{Bethe}}(\theta) = \sup_{\tau \in \mathbb{I}} \{ \langle \theta, \tau \rangle + H_{\text{Bethe}}(\tau) \} \]  (8)

\[ = \sup_{\tau \in \mathbb{I}} \left\{ \langle \theta, \tau \rangle + \sum_{v \in V(G)} H_v(\tau_v) - \sum_{(s,t) \in E(G)} I_{st}(\tau_{st}) \right\} \]  (9)

- Exact when \( G = T \) but we do this for any \( G \), still commutable
- we get an approximate log partition function, and approximate (pseudo) marginals (in \( \mathbb{I} \)), but this is perhaps much easier to compute.
- We can optimize this directly using a Lagrangian formulation.
Proposition 2.2

Consider a pairwise MRF with binary variables, with $A_{\text{Bethe}}(\theta)$ being the optimized free energy evaluated at a LBP fixed point $\tau = (\tau_s, s \in V; \tau_{st}, (s, t) \in E(G))$. Then we have the following relationship with the cumulant function $A(\theta)$.

$$A(\theta) = A_{\text{Bethe}}(\theta) + \log \left\{ 1 + \sum_{\emptyset \neq \tilde{E} \subseteq E} \beta_{\tilde{E}} \prod_{s \in V} \mathbb{E}_{\tau_s} \left[ (X_s - \tau_s)^{d_s(\tilde{E})} \right] \right\} \quad (10)$$

- Note that for any $\tilde{E}$ such that $\exists s$ with $d_s(\tilde{E}) = 1$, then the inner term is zero and vanishes.
Comparison of $A$ and $A_{\text{Bethe}}$

**Proposition 2.2**

Consider a pairwise MRF with binary variables, with $A_{\text{Bethe}}(\theta)$ being the optimized free energy evaluated at a LBP fixed point $
\tau = (\tau_s, s \in V; \tau_{st}, (s, t) \in E(G))$. Then we have the following relationship with the cumulant function $A(\theta)$.

\[
A(\theta) = A_{\text{Bethe}}(\theta) + \log \left\{ 1 + \sum_{\emptyset \neq \tilde{E} \subseteq E} \beta_{\tilde{E}} \prod_{s \in V} \mathbb{E}_{\tau_s} \left[ (X_s - \tau_s)^d_s(\tilde{E}) \right] \right\} \tag{10}
\]

- Note that for any $\tilde{E}$ such that $\exists s $ with $d_s(\tilde{E}) = 1$, then the inner term is zero and vanishes. **why?**
Proposition 2.2

Consider a pairwise MRF with binary variables, with $A_{\text{Bethe}}(\theta)$ being the optimized free energy evaluated at a LBP fixed point

$\tau = (\tau_s, s \in V; \tau_{st}, (s, t) \in E(G))$. Then we have the following relationship with the cumulant function $A(\theta)$.

$$A(\theta) = A_{\text{Bethe}}(\theta) + \log \left(1 + \sum_{\emptyset \neq \tilde{E} \subseteq E} \beta_{\tilde{E}} \prod_{s \in V} \mathbb{E}_{\tau_s} \left[(X_s - \tau_s)^{d_s(\tilde{E})}\right]\right)$$

(10)

- Note that for any $\tilde{E}$ such that $\exists s$ with $d_s(\tilde{E}) = 1$, then the inner term is zero and vanishes. Why? Thus, terms in the sum only exists for generalized loops.
Comparison of $A$ and $A_{\text{Bethe}}$

**Proposition 2.2**

Consider a pairwise MRF with binary variables, with $A_{\text{Bethe}}(\theta)$ being the optimized free energy evaluated at a LBP fixed point

\[
\tau = (\tau_s, s \in V; \tau_{st}, (s, t) \in E(G))
\]

Then we have the following relationship with the cumulant function $A(\theta)$.

\[
A(\theta) = A_{\text{Bethe}}(\theta) + \log \left\{ 1 + \sum_{\emptyset \neq \tilde{E} \subseteq E} \beta_{\tilde{E}} \prod_{s \in V} \mathbb{E}_{\tau_s} \left[ (X_s - \tau_s)^{d_s(\tilde{E})} \right] \right\} 
\]  

(10)

- Note that for any $\tilde{E}$ such that $\exists s$ with $d_s(\tilde{E}) = 1$, then the inner term is zero and vanishes. why? Thus, terms in the sum only exists for generalized loops.
- The generalized loops give the correction!
Consider a pairwise MRF with binary variables, with $A_{\text{Bethe}}(\theta)$ being the optimized free energy evaluated at a LBP fixed point $\tau = (\tau_s, s \in V; \tau_{st}, (s,t) \in E(G))$. Then we have the following relationship with the cumulant function $A(\theta)$.

$$A(\theta) = A_{\text{Bethe}}(\theta) + \log \left\{ 1 + \sum_{\emptyset \neq \tilde{E} \subseteq E} \beta_{\tilde{E}} \prod_{s \in V} \mathbb{E}_{\tau_s} \left[ (X_s - \tau_s)^{d_s(\tilde{E})} \right] \right\}$$  \hfill (10)

- Note that for any $\tilde{E}$ such that $\exists s$ with $d_s(\tilde{E}) = 1$, then the inner term is zero and vanishes. Why? Thus, terms in the sum only exists for generalized loops.
- The generalized loops give the correction!
- For trees, there are no generalized loops, and so if $G$ is a tree then we have an equality between $A(\theta)$ and $A_{\text{Bethe}}(\theta)$.
General idea of Kikuchi

- Variational representation of log partition function

\[ A(\theta) = \sup_{\mu \in \mathcal{M}} \{ \langle \theta, \mu \rangle - A^*(\mu) \} \quad \text{(11)} \]
General idea of Kikuchi

- Variational representation of log partition function

\[ A(\theta) = \sup_{\mu \in \mathcal{M}} \{ \langle \theta, \mu \rangle - A^*(\mu) \} \]  \hspace{1cm} (11)

- So far, we have used a replacement for \(-A^*(\mu)\) inspired by trees.
General idea of Kikuchi

- Variational representation of log partition function

\[
A(\theta) = \sup_{\mu \in \mathcal{M}} \{ \langle \theta, \mu \rangle - A^*(\mu) \} \tag{11}
\]

- So far, we have used a replacement for \(-A^*(\mu)\) inspired by trees.
- But we know a tree is really a 1-tree. Why not \(k\)-tree?
General idea of Kikuchi

- Variational representation of log partition function

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

(11)

- So far, we have used a replacement for \(-A^*(\mu)\) inspired by trees.

- But we know a tree is really a 1-tree. Why not \(k\)-tree?

- Why not some other junction tree?
General idea of Kikuchi

- Variational representation of log partition function

\[
A(\theta) = \sup_{\mu \in \mathcal{M}} \{ \langle \theta, \mu \rangle - A^*(\mu) \} \tag{11}
\]

- So far, we have used a replacement for \(-A^*(\mu)\) inspired by trees.
- But we know a tree is really a 1-tree. Why not \(k\)-tree?
- Why not some other junction tree?
- Junction trees are really hypertrees (special case of hypergraphs).
General idea of Kikuchi

- Variational representation of log partition function

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

(11)

- So far, we have used a replacement for \(-A^*(\mu)\) inspired by trees.
- But we know a tree is really a 1-tree. Why not \(k\)-tree?
- Why not some other junction tree?
- Junction trees are really hypertrees (special case of hypergraphs).
- So can we come up with:
General idea of Kikuchi

- Variational representation of log partition function

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

(11)

- So far, we have used a replacement for \(-A^*(\mu)\) inspired by trees.
- But we know a tree is really a 1-tree. Why not \(k\)-tree?
- Why not some other junction tree?
- Junction trees are really hypertrees (special case of hypergraphs).
- So can we come up with: 1) replacement for \(-A^*(\mu)\) associated with a hypertree/junction tree;
General idea of Kikuchi

- Variational representation of log partition function

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

(11)

- So far, we have used a replacement for \(-A^*(\mu)\) inspired by trees.
- But we know a tree is really a 1-tree. Why not \(k\)-tree?
- Why not some other junction tree?
- Junction trees are really hypertrees (special case of hypergraphs).
- So can we come up with: 1) replacement for \(-A^*(\mu)\) associated with a hypertree/junction tree; 2) a generalization for this replacement for any hypergraph; and
General idea of Kikuchi

- Variational representation of log partition function

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

(11)

- So far, we have used a replacement for \(-A^*(\mu)\) inspired by trees.
- But we know a tree is really a 1-tree. Why not \(k\)-tree?
- Why not some other junction tree?
- Junction trees are really hypertrees (special case of hypergraphs).
- So can we come up with: 1) replacement for \(-A^*(\mu)\) associated with a hypertree/junction tree; 2) a generalization for this replacement for any hypergraph; and 3) a corresponding generalized polytope associated with the hypergraph?
General idea of Kikuchi

- Variational representation of log partition function

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

(11)

- So far, we have used a replacement for \(-A^*(\mu)\) inspired by trees.
- But we know a tree is really a 1-tree. Why not \(k\)-tree?
- Why not some other junction tree?
- Junction trees are really hypertrees (special case of hypergraphs).
- So can we come up with: 1) replacement for \(-A^*(\mu)\) associated with a hypertree/junction tree; 2) a generalization for this replacement for any hypergraph; and 3) a corresponding generalized polytope associated with the hypergraph?

- This is the Kikuchi variational approach.
**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 as follows

$$
\log \varphi_h(x_h) \triangleq \sum_{g \preceq h} \omega(g, h) \log \mu_g(x_g) \tag{12}
$$

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

$$
\log \mu_h(x_h) = \sum_{g \preceq h} \log \varphi_g(x_g) \tag{13}
$$

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

$$
\mathcal{p}_\mu(x) = \prod_{h \in E} \varphi_h(x_h) \tag{14}
$$

this is the general way to factorize a distribution that factors w.r.t. a hypergraph, and when it is a 1-tree, we recover what we’ve already seen, i.e.
New expressions of entropy

- We can express entropic quantities as well, such as the hyperedge entropy

\[ H_h(\mu_h) = -\sum_{x_h} \mu_h(x_h) \log \mu_h(x_h) \]  \hspace{1cm} (15)

and the multi-information function

\[ I_h(\mu_h) = \sum_{x_h} \mu_h(x_h) \log \varphi_h(x_h) \]  \hspace{1cm} (16)
New expressions of entropy

- We can express entropic quantities as well, such as the hyperedge entropy

\[ H_h(\mu_h) = - \sum_{x_h} \mu_h(x_h) \log \mu_h(x_h) \]  \hspace{1cm} (15)

and the multi-information function

\[ I_h(\mu_h) = \sum_{x_h} \mu_h(x_h) \log \varphi_h(x_h) \]  \hspace{1cm} (16)

- In the case of a single tree edge \( h = (s, t) \), then \( I_h(\mu_h) = I(X_s; X_t) \) the standard mutual information.
New expressions of entropy

- We can express entropic quantities as well, such as the hyperedge entropy

\[
H_h(\mu_h) = - \sum_{x_h} \mu_h(x_h) \log \mu_h(x_h)
\]  

(15)

and the multi-information function

\[
I_h(\mu_h) = \sum_{x_h} \mu_h(x_h) \log \varphi_h(x_h)
\]  

(16)

- In the case of a single tree edge \( h = (s, t) \), then \( I_h(\mu_h) = I(X_s; X_t) \) the standard mutual information.

- Then the overall entropy of any hypertree distribution becomes

\[
H_{\text{hyper}}(\mu) = - \sum_{h \in E} I_h(\mu_h)
\]  

(17)
multi-information decomposition

- Using Möbius, we can write

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

(18)

(19)

where

\[
c(f) \equiv \sum_{e \geq f} \omega(f, e)
\]

(21)

This gives us a new expression for the hypertree entropy

\[
H_{\text{hyper}}(\mu) = \sum_{h \in E} c(h) H_h(\mu_h)
\]

(22)
Using Möbius, we can write

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

(18)

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

(19)

\[ = -\sum_{f \preceq h} c(f) H_f(\mu_f) \]  

(20)
Using Möbius, we can write

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

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

\[
= - \sum_{f \preceq h} c(f) H_f(\mu_f)
\]
multi-information decomposition

- Using Möbius, we can write

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

(18)

(19)

(20)

where

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

(21)
multi-information decomposition

- Using Möbius, we can write

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

(18)

\[ = \sum_{f \leq h} \sum_{e \geq f} \omega(e, f) \left\{ \sum_{x_f} \mu_f(x_f) \log \mu_f(x_f) \right\} \]  

(19)

\[ = - \sum_{f \leq h} c(f) H_f(\mu_f) \]  

(20)

where

\[ c(f) \triangleq \sum_{e \geq f} \omega(f, e) \]  

(21)

- This gives us a new expression for the hypertree entropy

\[ H_{\text{hyper}}(\mu) = \sum c(h) H_h(\mu_h) \]  

(22)
Usable to get Kikuchi variational approximation

- Given arbitrary hypergraph now, we’ll generalize the hypertree expressions above this arbitrary hypergraph, which will give us a variational expression that approximates cumulant.
Usable to get Kikuchi variational approximation

- Given arbitrary hypergraph now, we’ll generalize the hypertree expressions above this arbitrary hypergraph, which will give us a variational expression that approximates cumulant.

- Given hypergraph $G = (V, E)$, we have

$$p_\theta(x) \propto \exp \left\{ \sum_{h \in E} \sigma_h(x_h) \right\} \quad (23)$$

using same form of parameterization.
Usable to get Kikuchi variational approximation

- Given arbitrary hypergraph now, we’ll generalize the hypertree expressions above this arbitrary hypergraph, which will give us a variational expression that approximates cumulant.

- Given hypergraph \( G = (V, E) \), we have

\[
p_\theta(x) \propto \exp \left\{ \sum_{h \in E} \sigma_h(x_h) \right\}
\]

(23)

using same form of parameterization.

- Hypergraph will give us local marginal constraints on hypergraph pseudo marginals, i.e., for each \( h \in E \), we form marginal \( \tau_h(x_h) \) and define constraints

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

(24)
Usable to get Kikuchi variational approximation

- Sum to one constraint:

\[
\sum_{x_h} \tau_h(x_h) = 1
\]  
(25)
Usable to get Kikuchi variational approximation

- Sum to one constraint:

\[
\sum_{x_h} \tau_h(x_h) = 1 \quad (25)
\]

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

\[
\sum_{x_h \setminus g} \tau_h(x_h) = \tau_g(x_g) \quad (26)
\]
Usable to get Kikuchi variational approximation

- Sum to one constraint:
  \[
  \sum_{x_h} \tau_h(x_h) = 1 \tag{25}
  \]

- Local agreement via the hypergraph constraint. For any \( g \preceq h \) must have marginalization condition
  \[
  \sum_{x_h \setminus g} \tau_h(x_h) = \tau_g(x_g) \tag{26}
  \]

- Define new polyhedral constraint set \( \mathbb{I}_t(G) \)
  \[
  \mathbb{I}_t(G) = \{ \tau \geq 0 \mid \text{Equations (25) } \forall h, \text{ and (26) } \forall g \preceq h \text{ hold} \} \tag{27}
  \]
Kikuchi variational approximation

- Generalized entropy for the hypergraph:

\[
H_{\text{app}} = \sum_{g \in E} c(g) H_g(\tau_g) \tag{28}
\]

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

\[
c(g) = \sum_{f \supseteq g} \omega(g, f) \tag{29}
\]
Kikuchi variational approximation

- Generalized entropy for the hypergraph:

\[ H_{\text{app}} = \sum_{g \in E} c(g) H_g(\tau_g) \]  \hspace{1cm} (28)

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

\[ c(g) = \sum_{f \succeq g} \omega(g, f) \]  \hspace{1cm} (29)

- This at last gets the Kikuchi variational approximation

\[ A_{\text{Kikuchi}}(\theta) = \max_{\tau \in \mathbb{L}_t(G)} \left\{ \langle \theta, \tau \rangle + H_{\text{app}}(\tau) \right\} \]  \hspace{1cm} (30)

For a graph, this is exactly \( A_{\text{Bethe}}(\theta) \). If, on the other hand, the graph is a junction tree, then this is exact (although it might be expensive, exponential in the tree-width to compute \( H_{\text{app}} \)).

Can define message passing algorithms on the hypertree, and show that if it converges, it is a fixed point of the Lagrangian associated with this.
Kikuchi variational approximation

- Generalized entropy for the hypergraph:

\[ H_{\text{app}} = \sum_{g \in E} c(g) H_g(\tau_g) \]  \hspace{1cm} (28)

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

\[ c(g) = \sum_{f \succeq g} \omega(g,f) \]  \hspace{1cm} (29)

- This at last gets the Kikuchi variational approximation

\[ A_{\text{Kikuchi}}(\theta) = \max_{\tau \in \mathbb{L}_t(G)} \{ \langle \theta, \tau \rangle + H_{\text{app}}(\tau) \} \]  \hspace{1cm} (30)

- For a graph, this is exactly \( A_{\text{Bethe}}(\theta) \). If, on the other hand, the graph is a junction tree, then this is exact (although it might be expensive, exponential in the tree-width to compute \( H_{\text{app}} \)).
Kikuchi variational approximation

- Generalized entropy for the hypergraph:

\[ H_{app} = \sum_{g \in E} c(g) H_g(\tau_g) \] (28)

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

\[ c(g) = \sum_{f \succeq g} \omega(g, f) \] (29)

- This at last gets the Kikuchi variational approximation

\[ A_{Kikuchi}(\theta) = \max_{\tau \in \mathbb{L}_t(G)} \{ \langle \theta, \tau \rangle + H_{app}(\tau) \} \] (30)

- For a graph, this is exactly \( A_{Bethe}(\theta) \). If, on the other hand, the graph is a junction tree, then this is exact (although it might be expensive, exponential in the tree-width to compute \( H_{app} \)).

- Can define message passing algorithms on the hypertree, and show that if it converges, it is a fixed point of the Lagrangian associated with this.
Conjugate Duality

**Theorem 3.1 (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:

\[
A^*(\mu) = \sup_{\theta \in \Omega} (\langle \theta, \mu \rangle - A(\theta)) = \begin{cases} 
-H(p_{\theta(\mu)}) & \text{if } \mu \in \mathcal{M}^\circ \\
+\infty & \text{if } \mu \in \bar{\mathcal{M}}
\end{cases}
\] (1)

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

\[
A(\theta) = \sup_{\mu \in \mathcal{M}} \{ \langle \theta, \mu \rangle - A^*(\mu) \} 
\] (2)

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

\[
\mu = \int_{\mathcal{D}_X} \phi(x)p_{\theta}(x)\nu(dx) = \mathbb{E}_{\theta}[\phi(X)] = \nabla A(\theta)
\] (3)
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 2.
Term Decoupling

- Partition the sufficient statistics into two parts:
Term Decoupling

- Partition the sufficient statistics into two parts:
  - Tractable component

\[ \phi \triangleq (\phi_1, \phi_2, \ldots, \phi_{d_T}) \]  \hspace{1cm} (31)
Term Decoupling

- Partition the sufficient statistics into two parts:
  - Tractable component
    \[ \phi \triangleq (\phi_1, \phi_2, \ldots, \phi_{d_T}) \] (31)
  - Intractable component
    \[ \Phi \triangleq (\Phi^1, \Phi^2, \ldots, \Phi^{d_I}) \] (32)
Term Decoupling

- Partition the sufficient statistics into two parts:
  - Tractable component
    \[ \phi \triangleq (\phi_1, \phi_2, \ldots, \phi_{d_T}) \] (31)
  - Intractable component
    \[ \Phi \triangleq (\Phi^1, \Phi^2, \ldots, \Phi^{d_I}) \] (32)
- \( \phi_i \) are typically univariate, while \( \Phi^i \) are multivariate \((b\text{-dimensional})\).
Term Decoupling

- Partition the sufficient statistics into two parts:
  - Tractable component
    \[
    \phi \triangleq (\phi_1, \phi_2, \ldots, \phi_{d_T}) \tag{31} \]
  - Intractable component
    \[
    \Phi \triangleq (\Phi^1, \Phi^2, \ldots, \Phi^{d_I}) \tag{32} \]

- \(\phi_i\) are typically univariate, while \(\Phi^i\) are multivariate (\(b\)-dimensional).
- Consider exponential families associated with subcollection \((\phi, \Phi)\).
Tractable component

\( \phi \triangleq (\phi_1, \phi_2, \ldots, \phi_{dT}) \) (33)
Tractable component

$\phi \triangleq (\phi_1, \phi_2, \ldots, \phi_{d_T})$ (33)

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

- Tractable component

\[ \phi \triangleq (\phi_1, \phi_2, \ldots, \phi_{d_T}) \] (33)

- 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
Intractable component

\[ \Phi \triangleq (\Phi_1, \Phi_2, \ldots, \Phi_{d_I}) \] (34)
Intractable component

- Intractable component

\[ \Phi \triangleq (\Phi_1, \Phi_2, \ldots, \Phi_{d_I}) \]  \hspace{1cm} (34)

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

- Intractable component

\[ \Phi \triangleq (\Phi_1, \Phi_2, \ldots, \Phi_{d_I}) \quad (34) \]

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

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

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

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

- The associated exponential family

\[ p(x; \theta, \tilde{\theta}) \propto f_0(x) \exp \left( \langle \theta, \phi(x) \rangle \right) \exp \left( \langle \tilde{\theta}, \Phi(x) \rangle \right) \]  
\[ = f_0(x) \exp \left( \langle \theta, \phi(x) \rangle \right) \prod_{i=1}^{d_I} \exp \left( \langle \tilde{\theta}^i, \Phi^i(x) \rangle \right) \]
Associated Distributions

- The associated exponential family

\[ p(x; \theta, \tilde{\theta}) \propto f_0(x) \exp \left( \langle \theta, \phi(x) \rangle \right) \exp \left( \left\langle \tilde{\theta}, \Phi(x) \right\rangle \right) \]  

\[ = f_0(x) \exp \left( \langle \theta, \phi(x) \rangle \right) \prod_{i=1}^{d_I} \exp \left( \left\langle \tilde{\theta}^i, \Phi^i(x) \right\rangle \right) \]  

- Base model is tractable

\[ p(x; \theta, \tilde{0}) \propto f_0(x) \exp \left( \langle \theta, \phi(x) \rangle \right) \]
Associated Distributions

- The associated exponential family

\[
p(x; \theta, \tilde{\theta}) \propto f_0(x) \exp (\langle \theta, \phi(x) \rangle) \exp \left( \langle \tilde{\theta}, \Phi(x) \rangle \right)
\]

\[
= f_0(x) \exp (\langle \theta, \phi(x) \rangle) \prod_{i=1}^{d_I} \exp \left( \langle \tilde{\theta}^i, \Phi^i(x) \rangle \right)
\]

- Base model is tractable

\[
p(x; \theta, \tilde{\theta}) \propto f_0(x) \exp (\langle \theta, \phi(x) \rangle)
\]

- \(\Phi^i\)-augmented model

\[
p(x; \theta, \tilde{\theta}^i) \propto f_0(x) \exp (\langle \theta, \phi(x) \rangle) \exp \left( \langle \tilde{\theta}^i, \Phi^i(x) \rangle \right)
\]
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).
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 $\varphi(y; \mu, \Lambda)$ be Normal with mean $\mu$ covariance $\Lambda$. 
Example: Mixture models

- Let $\varphi(y; \mu, \Lambda)$ be Normal with mean $\mu$ covariance $\Lambda$.
- Two component Gaussian mixture model

$$p(y|X = x) = (1 - \alpha)\varphi(y; 0, \sigma_0^2 I) + \alpha\varphi(y; x, \sigma_1^2 I) \quad (39)$$
Example: Mixture models

- Let $\varphi(y; \mu, \Lambda)$ be Normal with mean $\mu$ covariance $\Lambda$.
- Two component Gaussian mixture model

$$p(y|X = x) = (1 - \alpha)\varphi(y; 0, \sigma_0^2I) + \alpha\varphi(y; x, \sigma_1^2I)$$  \hspace{1cm} (39)

- Assume $n$ 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 $\varphi(y; \mu, \Lambda)$ be Normal with mean $\mu$ covariance $\Lambda$.
- Two component Gaussian mixture model

\[
p(y|X = x) = (1 - \alpha)\varphi(y; 0, \sigma_0^2 I) + \alpha\varphi(y; x, \sigma_1^2 I) \tag{39}\]

- Assume $n$ 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

\[
p(x|y^1, \ldots, y^n) \propto \exp \left( -\frac{1}{2} x^\top \sigma^{-1} x \right) \prod_{i=1}^n p(y^i|X = x) \tag{40}\]

\[
= \exp \left( -\frac{1}{2} x^\top \sigma^{-1} x \right) \exp \left\{ \sum_{i=1}^n \log p(y^i|X = x) \right\} \tag{41}\]
Example: Mixture models

- We equate \( \exp \left(-\frac{1}{2} x^\top \sigma^{-1} x \right) \) with \( f_0(x) \exp(\langle \theta, \phi(x) \rangle) \), with \( d_T = m \).
Example: Mixture models

- We equate $\exp \left( -\frac{1}{2} x^\top \sigma^{-1} x \right)$ with $f_0(x) \exp(\langle \theta, \phi(x) \rangle)$, with $d_T = m$.
- Such a distribution is Gaussian, and getting marginals is “cheap” $O(m^3)$.
Example: Mixture models

- We equate \(\exp\left(-\frac{1}{2}x^\top \sigma^{-1} x\right)\) with \(f_0(x) \exp(\langle \theta, \phi(x) \rangle)\), with \(d_T = m\).
- Such a distribution is Gaussian, and getting marginals is “cheap” \(O(m^3)\).
- Also, \(\exp \left\{ \sum_{i=1}^n \log p(y^i|X = x) \right\} \) cor. to \(\prod_{i=1}^{d_I} \exp \left(\langle \tilde{\theta}^i, \Phi^i(x) \rangle\right)\), with \(b = 1\).
Example: Mixture models

- We equate $\exp\left(-\frac{1}{2}x^\top \sigma^{-1}x\right)$ with $f_0(x) \exp(\langle \theta, \phi(x) \rangle)$, with $d_T = m$.
- Such a distribution is Gaussian, and getting marginals is “cheap” $O(m^3)$.
- Also, $\exp \left\{ \sum_{i=1}^{n} \log p(y^i|X = x) \right\}$ cor. to $\prod_{i=1}^{d_I} \exp \left( \langle \tilde{\theta}^i, \Phi^i(x) \rangle \right)$, with $b = 1$.
- Base distribution $p(x; \theta, \tilde{\theta}) \propto \exp\left(-\frac{1}{2}x^\top \sigma^{-1}x\right)$ which is a Gaussian and easy.
Example: Mixture models

- We equate $\exp\left(-\frac{1}{2}x^\top \sigma^{-1} x\right)$ with $f_0(x) \exp(\langle \theta, \phi(x) \rangle)$, with $dT = m$.
- Such a distribution is Gaussian, and getting marginals is “cheap” $O(m^3)$.
- Also, $\exp \left\{ \sum_{i=1}^{n} \log p(y^i|X = x) \right\}$ cor. to $\prod_{i=1}^{dI} \exp \left( \langle \tilde{\theta}^i, \Phi^i(x) \rangle \right)$, with $b = 1$.
- Base distribution $p(x; \theta, \tilde{\theta}) \propto \exp \left(-\frac{1}{2}x^\top \sigma^{-1} x\right)$ which is a Gaussian and easy.
- If we multiply in one intractable term, 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 \( f_0(x) \exp(\langle \theta, \phi(x) \rangle) \), with \( d_T = m \).
- Such a distribution is Gaussian, and getting marginals is “cheap” \( O(m^3) \).
- Also, \( \exp \left\{ \sum_{i=1}^{n} \log p(y^i|X = x) \right\} \) cor. to \( \prod_{i=1}^{d_I} \exp \left( \langle \tilde{\theta}^i, \Phi^i(x) \rangle \right) \), with \( b = 1 \).
- Base distribution \( p(x; \theta, \tilde{\theta}) \propto \exp \left( -\frac{1}{2} x^\top \sigma^{-1} x \right) \) which is a Gaussian and easy.
- If we multiply in one intractable term, still not so bad (quite easy in fact).
- I.e., \( \Phi^i \)-augmented distribution is proportional to

\[
\exp \left( -\frac{1}{2} x^\top \sigma^{-1} x \right) \left[ (1 - \alpha) \varphi(y^i; 0, \sigma_0^2 I) + \alpha \varphi(y^i; x, \sigma_1^2 I) \right] \quad (42)
\]
Example: Mixture models

- We equate \( \exp\left(-\frac{1}{2}x^\top \sigma^{-1} x\right) \) with \( f_0(x) \exp(\langle \theta, \phi(x) \rangle) \), with \( d_T = m \).
- Such a distribution is Gaussian, and getting marginals is “cheap” \( O(m^3) \).
- Also, \( \exp \left\{ \sum_{i=1}^n \log p(y^i|X = x) \right\} \) cor. to \( \prod_{i=1}^{d_I} \exp \left(\langle \tilde{\theta}^i, \Phi^i(x) \rangle\right) \), with \( b = 1 \).
- 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.
- If we multiply in one intractable term, still not so bad (quite easy in fact).
- I.e., \( \Phi^i \)-augmented distribution is proportional to
  \[
  \exp\left(-\frac{1}{2}x^\top \sigma^{-1} x\right) \left[(1 - \alpha)\varphi(y^i; 0, \sigma_0^2 I) + \alpha \varphi(y^i; x, \sigma_1^2 I)\right]
  \]
- Computing marginals on this is not too bad (mixture of 2 components)
Example: Mixture models

- We equate $\exp\left(-\frac{1}{2} x^\top \sigma^{-1} x\right)$ with $f_0(x) \exp(\langle \theta, \phi(x) \rangle)$, with $d_T = m$.
- Such a distribution is Gaussian, and getting marginals is “cheap” $O(m^3)$.
- Also, $\exp\left\{ \sum_{i=1}^{n} \log p(y_i|X = x) \right\}$ cor. to $\prod_{i=1}^{d_I} \exp \left( \langle \tilde{\theta}^i, \Phi^i(x) \rangle \right)$, with $b = 1$.
- 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.
- If we multiply in one intractable term, still not so bad (quite easy in fact).
- I.e., $\Phi^i$-augmented distribution is proportional to

$$
\exp \left(-\frac{1}{2} x^\top \sigma^{-1} x\right) \left[ (1 - \alpha) \phi(y^i; 0, \sigma_0^2 I) + \alpha \phi(y^i; x, \sigma_1^2 I) \right] \tag{42}
$$

- Computing marginals on this is not too bad (mixture of 2 components)
- If we multiply in all $\Phi^i$, becomes intractable ($2^n$ components)
Polytope and Base case

- We can partition the mean parameters \((\mu, \tilde{\mu}) \in \mathbb{R}^{dT+dI \times b}\)
Polytope and Base case

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

- Marginal polytope associated with these means

\[
\mathcal{M}(\phi, \Phi) = \{(\mu, \tilde{\mu})|(\mu, \tilde{\mu}) = \mathbb{E}_p[(\phi(X), \Phi(X))]| \text{ for some } p\} \quad (43)
\]

along with negative dual of cumulant, or entropy \(H(\mu, \tilde{\mu}) = -A^*(\mu, \tilde{\mu})\).
We can partition the mean parameters $(\mu, \tilde{\mu}) \in \mathbb{R}^{dT+dI \times b}$

Marginal polytope associated with these means

\[ \mathcal{M}(\phi, \Phi) = \{ (\mu, \tilde{\mu}) | (\mu, \tilde{\mu}) = \mathbb{E}_p[(\phi(X), \Phi(X))] \} \text{ for some } p \] (43)

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

\[ \mathcal{M}(\phi) = \left\{ \mu \in \mathbb{R}^{dT} | \mu = \mathbb{E}_p(\phi(X)) \right\} \] (44)
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

$$M(\phi, \Phi) = \{(\mu, \tilde{\mu}) | (\mu, \tilde{\mu}) = \mathbb{E}_p[(\phi(X), \Phi(X))] \text{ for some } p\} \quad (43)$$

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

$$M(\phi) = \left\{ \mu \in \mathbb{R}^{d_T} | \mu = \mathbb{E}_p(\phi(X)) \right\} \quad (44)$$

- 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

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

(45)
For each $i = 1 \ldots d_I$ we have a $\Phi^i$-augmented exp. model and polytope

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

(45)

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

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

(45)

- 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).
Augmented Base case

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

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

(45)

- 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 \((\tau, \tilde{\tau})\) define “projection operation”

\[
\Pi^i(\tau, \tilde{\tau}) \rightarrow (\tau, \tilde{\tau}^i)
\] (46)
New outer bound

- For any mean parms \((\tau, \tilde{\tau})\) define “projection operation”

\[
\Pi^i(\tau, \tilde{\tau}) \rightarrow (\tau, \tilde{\tau}^i)
\]  \hspace{1cm} (46)

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

\[
\mathcal{L}(\phi, \Phi) = \left\{ (\tau, \tilde{\tau}) | \tau \in M(\phi), \Pi^i(\tau, \tilde{\tau}) \in M(\phi, \Phi^i), \forall i \right\}
\]  \hspace{1cm} (47)
New outer bound

- For any mean parms \((\tau, \tilde{\tau})\) define “projection operation”

\[
\Pi^i(\tau, \tilde{\tau}) \rightarrow (\tau, \tilde{\tau}^i)
\]  \hspace{1cm} (46)

- 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} (47)

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

- For any mean parms \((\tau, \tilde{\tau})\) define “projection operation”

\[
\Pi^i(\tau, \tilde{\tau}) \rightarrow (\tau, \tilde{\tau}^i)
\] (46)

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

\[
\mathcal{L}(\phi, \Phi) = \{ (\tau, \tilde{\tau}) | \tau \in M(\phi), \Pi^i(\tau, \tilde{\tau}) \in M(\phi, \Phi^i), \forall i \}
\] (47)

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

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

For any mean parms \((\tau, \tilde{\tau}) \in L(\phi, \Phi)\)

- There is a member of the \(\phi\)-exponential family which mean parameters \(\tau\) with entropy \(H(\tau)\)
Members in new outer bound

For any mean params \((\tau, \tilde{\tau}) \in \mathcal{L}(\phi, \Phi)\):

- There is a member of the \(\phi\)-exponential family which mean parameters \(\tau\) with entropy \(H(\tau)\).
- 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 \((\tau, \tilde{\tau}) \in \mathcal{L}(\phi, \Phi)\)

- There is a member of the \(\phi\)-exponential family which mean parameters \(\tau\) with entropy \(H(\tau)\)
- 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 \(H(\tau, \tilde{\tau}^i)\)
- Consider new entropy approximation

\[
H(\tau, \tilde{\tau}) \approx H_{ep}(\tau, \tilde{\tau}) \triangleq H(\tau) + \sum_{\ell=1}^{d_I} \left[ H(\tau, \tilde{\tau}^\ell) - H(\tau) \right] \quad (48)
\]
Members in new outer bound

For any mean parms \((\tau, \tilde{\tau}) \in \mathcal{L}(\phi, \Phi)\)

- There is a member of the \(\phi\)-exponential family which mean parameters \(\tau\) with entropy \(H(\tau)\)
- 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 \(H(\tau, \tilde{\tau}^i)\)
- Consider new entropy approximation

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

(48)

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

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

(49)
Members in new outer bound

For any mean params \((\tau, \tilde{\tau}) \in \mathcal{L}(\phi, \Phi)\)

- There is a member of the \(\phi\)-exponential family which mean parameters \(\tau\) with entropy \(H(\tau)\)
- 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 \(H(\tau, \tilde{\tau}^i)\)
- Consider new entropy approximation

\[
H(\tau, \tilde{\tau}) \approx H_{\text{ep}}(\tau, \tilde{\tau}) \triangleq H(\tau) + \sum_{\ell=1}^{d_I} \left[ H(\tau, \tilde{\tau}^\ell) - H(\tau) \right] \tag{48}
\]

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

\[
\max_{(\tau, \tilde{\tau}) \in \mathcal{L}(\phi, \Phi)} \left\{ \langle \tau, \theta \rangle + \langle \tilde{\tau}, \tilde{\theta} \rangle + H_{\text{ep}}(\tau, \tilde{\tau}) \right\} \tag{49}
\]

- This covers the EP algorithms.
Members in new outer bound

For any mean parms \((\tau, \tilde{\tau}) \in \mathcal{L}(\phi, \Phi)\)

- There is a member of the \(\phi\)-exponential family which mean parameters \(\tau\) with entropy \(H(\tau)\)
- 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 \(H(\tau, \tilde{\tau}^i)\)
- Consider new entropy approximation

\[
H(\tau, \tilde{\tau}) \approx H_{ep}(\tau, \tilde{\tau}) \triangleq H(\tau) + \sum_{\ell=1}^{d_I} \left[ H(\tau, \tilde{\tau}^l) - H(\tau) \right]
\] (48)

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

\[
\max_{(\tau, \tilde{\tau}) \in \mathcal{L}(\phi, \Phi)} \left\{ \langle \tau, \theta \rangle + \langle \tilde{\tau}, \tilde{\theta} \rangle + H_{ep}(\tau, \tilde{\tau}) \right\}
\] (49)

- This covers the EP algorithms.
- When we take \(\phi\) to be unaries and \(\Phi\) to be edges, we get exactly Bethe approximation.
Moment Matching

- We may define a Lagrangian version of the objective

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

(50)

where

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

(51)
Moment Matching

- We may define a Lagrangian version of the objective

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

(50)

where

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

(51)

- Resulting exponential family may be written as:

\[ q(x; \theta, \lambda) \propto f_0(x) \exp \left\{ \left\langle \theta + \sum_{i=1}^{d_I} \lambda^i, \phi(x) \right\rangle \right\} \]  

(52)
Moment Matching $\rightarrow$ Expectation Propagation Updates

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

For the following distribution $q_i(x; \theta, \tilde{\theta}_i, \lambda)$

$$q_i(x; \theta, \tilde{\theta}_i, \lambda) \propto f_0(x) \exp \left[ \langle \theta + \sum_{\ell \neq i} \lambda_{\ell}, \phi(x) \rangle + \langle \tilde{\theta}_i, \Phi_i(x) \rangle \right]$$

and then compute the mean parameters $\eta_i(n) = \int q_i(n)(x) \phi(x) \nu(dx) = \mathbb{E}_{q_i(n)}[\phi(X)]$.
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. For the following distribution

\[
q^i(x; \theta, \tilde{\theta}^i, \lambda) \propto f_0(x) \exp \left( \langle \theta + \sum_{\ell \neq i} \lambda^\ell, \phi(x) \rangle + \langle \tilde{\theta}^i, \Phi^i(x) \rangle \right) \tag{53}
\]

and then compute the mean parameters

\[
\eta^{i(n)} = \int q^{i(n)}(x) \phi(x) \nu(dx) = \mathbb{E}_{q^{i(n)}}[\phi(X)] \tag{54}
\]
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. For the following distribution

$$q^i(x; \theta, \tilde{\theta}^i, \lambda) \propto f_0(x) \exp \left( \left\langle \theta + \sum_{\ell \neq i} \lambda^\ell, \phi(x) \right\rangle + \left\langle \tilde{\theta}^i, \Phi^i(x) \right\rangle \right)$$

and then compute the mean parameters

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

4. Form base distribution $q$ using Equation 52 and adjust $\lambda^{i(n)}$ to satisfy the moment-matching condition

$$\mathbb{E}_{q}[\phi(X)] = \eta^{i(n)}$$
When base distribution is unaries and \( \Phi^i \) is the edges of a graph, we get Bethe approximation, and standard sum-product LBP.
Moment Matching $\rightarrow$ Expectation Propagation Updates

1. When base distribution is unaries and $\Phi_i$ is the edges of a graph, we get Bethe approximation, and standard sum-product LBP.

2. When base distribution is a tree, we get tree-structured EP
Moment Matching $\rightarrow$ Expectation Propagation Updates

1. When base distribution is unaries and $\Phi^i$ is the edges of a graph, we get Bethe approximation, and standard sum-product LBP.
2. When base distribution is a tree, we get tree-structured EP
3. Can also be done for Gaussian mixture models. More details in text.
Most of this material comes from the Wainwright and Jordan book.