EE512A – Advanced Inference in Graphical Models

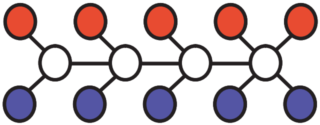
— Fall Quarter, Lecture 16 —

http://j.ee.washington.edu/~bilmes/classes/ee512a_fall_2014/

Prof. Jeff Bilmes

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

Nov 24th, 2014



Announcements

- Wainwright and Jordan *Graphical Models, Exponential Families, and Variational Inference* `http://www.nowpublishers.com/product.aspx?product=MAL&doi=2200000001`
- 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).

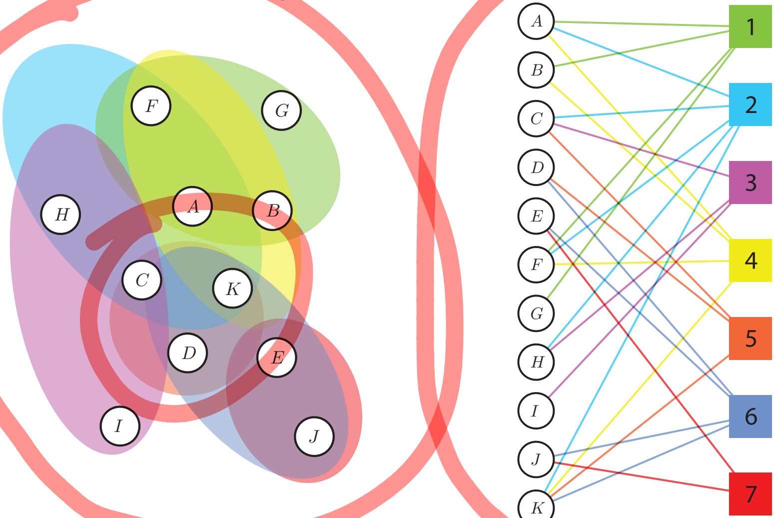
Class Road Map - EE512a

- 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.

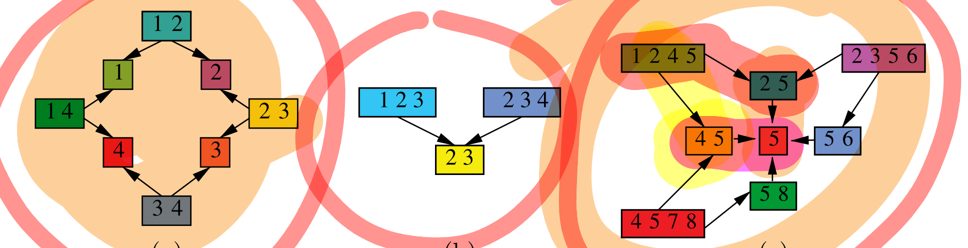
Drawing/Visualizing Hypergraphs as Bipartite Graphs

- 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.



(a) (b) (c)

- 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.

$$\forall A \subseteq V : \Upsilon(A) = \sum_{B : B \subseteq A} \Omega(B) \qquad (16.13)$$

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

- 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 $\preceq$.
- A zeta function of a poset is a mapping $\zeta : \mathcal{P} \times \mathcal{P} \to \mathbb{R}$ defined by

$$\zeta(g,h) = \begin{cases} 1 & \text{if } g \preceq h, \\ 0 & \text{otherwise.} \end{cases} \tag{16.23}$$

- 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 $h : h \not\preceq g$.
- Given $\omega(g,f)$ defined for f such that $g \preceq f \prec h$, we define

$$\omega(g,h) = - \sum_{\{f | g \preceq f \prec h\}} \omega(g,f) \tag{16.24}$$

- Then, ω and ζ are multiplicative inverses, in that

$$\sum_{f \in \mathcal{P}} \omega(g,f)\zeta(f,h) = \sum_{\{f | g \preceq f \preceq h\}} \omega(g,f) = \delta(g,h) \tag{16.25}$$

General Möbius Inversion Lemma for Posets

Lemma 16.2.8 (General Möbius Inversion Lemma)

Given real valued functions Υ and Ω defined on poset $\mathcal{P}$, then $\Omega(h)$ may be expressed via $\Upsilon(\cdot)$ via

$$\Omega(h) = \sum_{g \preceq h} \Upsilon(g) \quad \textit{for all } h \in \mathcal{P} \tag{16.23}$$

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

$$\Upsilon(h) = \sum_{g \preceq h} \Omega(g)\omega(g,h) \quad \textit{for all } h \in \mathcal{P} \tag{16.24}$$

When $\mathcal{P} = 2^V$ 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

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

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

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

- Key, when φ_h is defined as above, and G is a hypertree we have

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

$\Rightarrow$ 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)

$$c(f) \triangleq \sum_{e \succeq f} \omega(f,e) \tag{16.31}$$

- This gives us a new expression for the hypertree entropy

$$H_{\text{hyper}}(\mu) = \sum_{h \in E} c(h) H_h(\mu_h) \tag{16.32}$$

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 $\mathbb{L}_t(G)$

$$\mathbb{L}_t(G) = \{\tau \geq 0 | \text{ Equations (16.47) } \forall h, \text{ and (16.55) } \forall g \preceq h \text{ hold}\} \tag{16.35}$$

Kikuchi variational approximation, entropy approx

- Generalized approximate (app) entropy for the hypergraph:

$$H_{\text{app}} = \sum_{g \in E} c(g) H_g(\tau_g) \tag{16.33}$$

where H_g is hyperedge entropy and overcounting number defined by:

$$c(g) = \sum_{f \succeq g} \omega(g, f) \tag{16.34}$$

Variational Approach Amenable to Approximation

- Original variational representation of log partition function

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

where dual takes form:

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

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:

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

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

Variational Approach Amenable to Approximation

- Original variational representation of log partition function

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

where dual takes form:

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

- 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 $\mathcal{M}$ 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

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

Kikuchi variational approximation

- This at last gets the Kikuchi variational approximation

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

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

Kikuchi variational approximation

- This at last gets the Kikuchi variational approximation

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

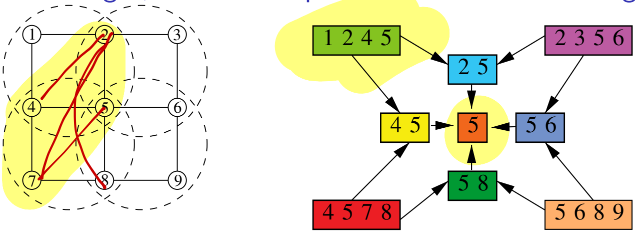
- For a graph, this is exactly $A_{\mathsf{Bethe}}(\theta)$.
- Also, if hypergraph is junction tree (r.i.p. holds, tree-local consistency implies global consistency), then also exact (although expensive, exponential in the tree-width to compute H_{app}).

Kikuchi variational approximation

- This at last gets the Kikuchi variational approximation

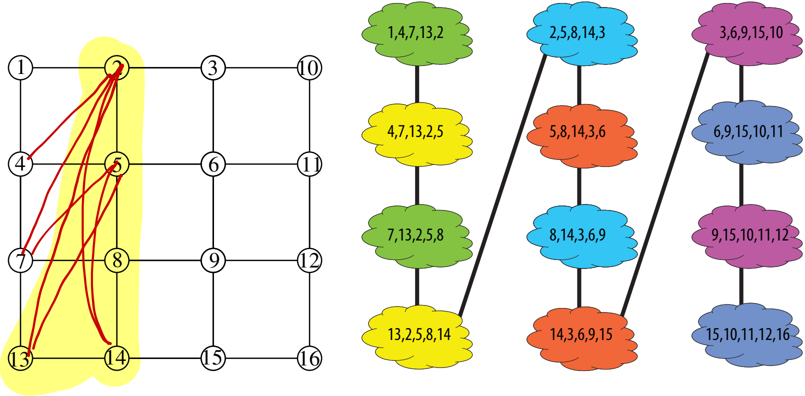
$$A_{\text{Kikuchi}}(\theta) = \max_{\tau \in \mathbb{L}_t(G)} \{\langle \theta, \tau \rangle + H_{\text{app}}(\tau)\} \tag{16.1}$$

- 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_{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.



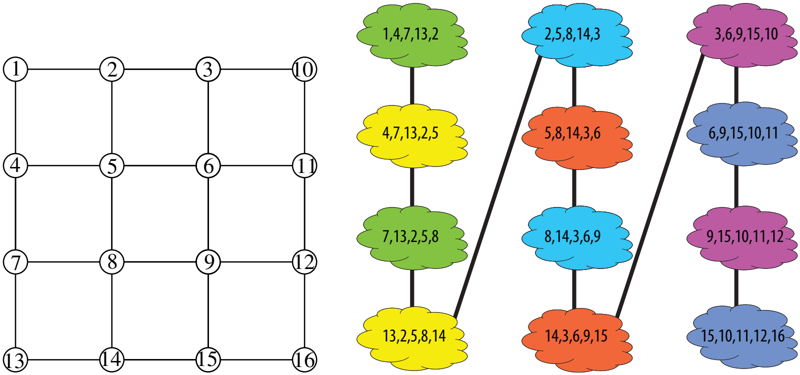
Kikuchi variational approximation, 3x3 grid example

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



Kikuchi variational approximation, 3x3 grid example

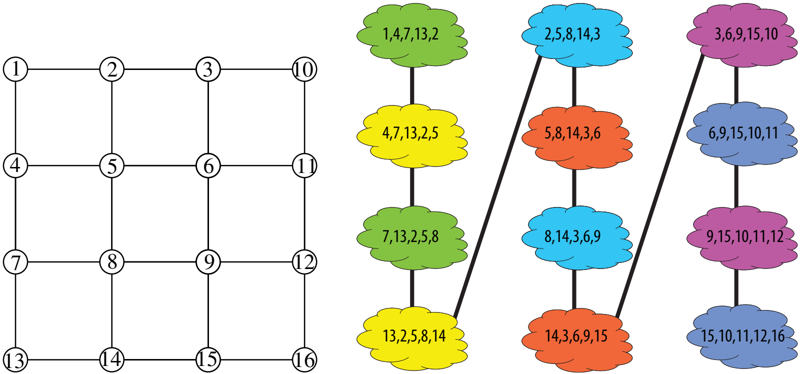
- Example, left is 3x3 grid, right is optimal junction tree cover.



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

Kikuchi variational approximation, 3x3 grid example

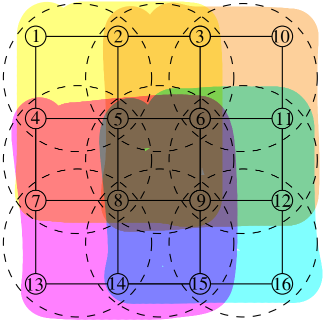
- Example, left is 3x3 grid, right is optimal junction tree cover.

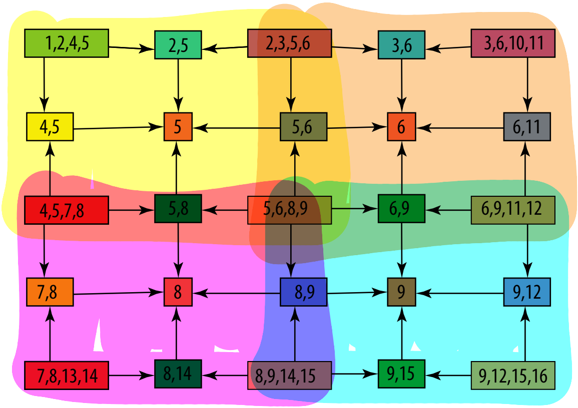


- Treewidth is 4, so complexity is $O(r^5)$.
- In general, for $n \times n$ grid strutured graph, treewidth is $O(n)$ (grows as the square root of the number of nodes).

Kikuchi variational approximation, 3x3 grid example

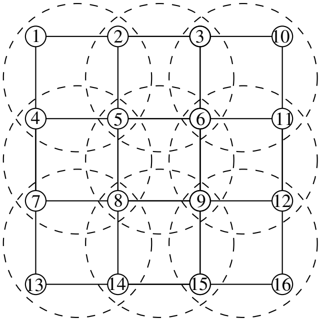
- Left is clustering of vertices in 3x3 grid, and right is hyperedge graph/region graph.

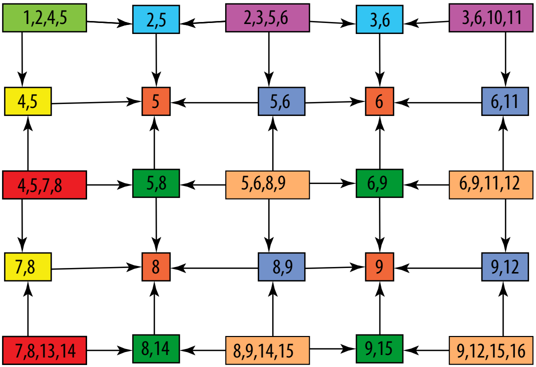




Kikuchi variational approximation, 3x3 grid example

- 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 ndoe sets)

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 ndoe 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 ndoe 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 ndoe 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.

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 ndoe 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!

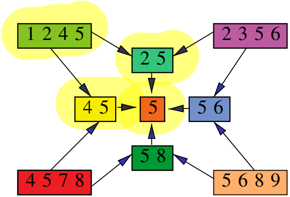
Generalized BP (GBP): Key idea

- 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 ndoe 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

$$A_{\text{Kikuchi}}(\theta) = \max_{\tau \in \mathbb{L}_t(G)} \{\langle \theta, \tau \rangle + H_{\text{app}}(\tau)\} \tag{16.2}$$

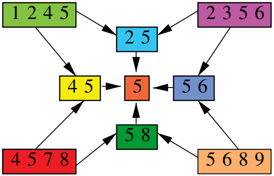
GBP examples: parent-to-child

- 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 $\{4, 5\}$.



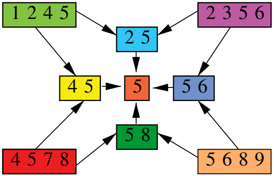
GBP examples: parent-to-child

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

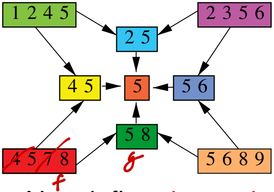
GBP examples: parent-to-child



- 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 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 $\{4, 5\}$.

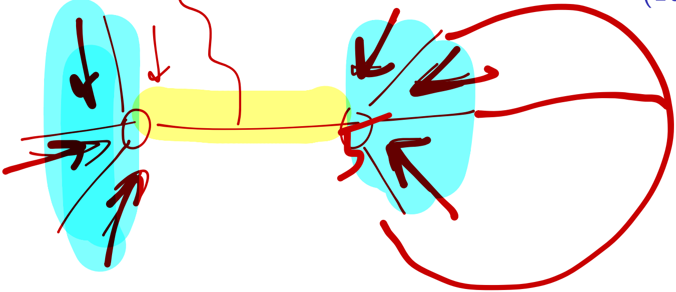


- For $h \in E$, let Par(h) be the set of parents. Also define descendants as $\mathcal{D}(h) = \{g \in E | g \prec h\}$ and ancestors as $\mathcal{A}(h) = \{g \in E | g \succ h\}$.
- Also define $\mathcal{D}^+(h) = \mathcal{D}(h) \cup \{h\}$ and $\mathcal{A}^+(h) = \mathcal{A}(h) \cup \{h\}$
- If $f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

GBP examples: parent-to-child message

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

$$\tau_h(x_h) \propto \left[\prod_{g \in \mathcal{D}^+(h)} \exp(\theta(x_g)) \right] \left[\prod_{g \in \mathcal{D}^+(h)} \prod_{f \in \mathrm{Par}(g) \setminus \mathcal{D}^+(h)} M_{f \to g}(x_g) \right] \tag{16.3}$$

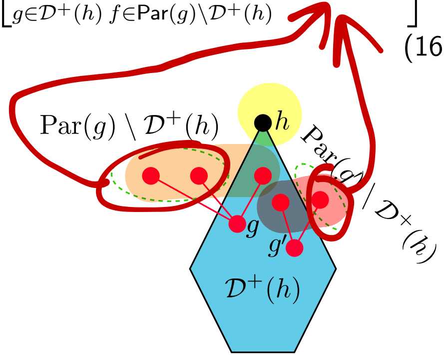


GBP examples: parent-to-child message

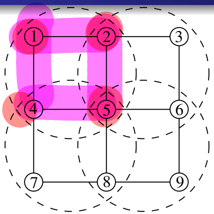
- Then parent-to-child message passing takes the form:

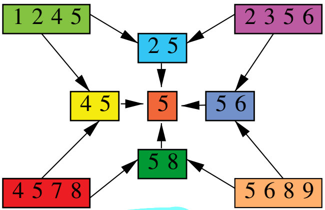
$$\tau_h(x_h) \propto \left[\prod_{g \in \mathcal{D}^+(h)} \exp(\theta(x_g)) \right] \left[\prod_{g \in \mathcal{D}^+(h)} \prod_{f \in \mathsf{Par}(g) \setminus \mathcal{D}^+(h)} M_{f \to g}(x_g) \right] \tag{16.3}$$

- 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.



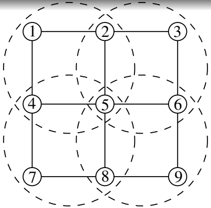
GBP examples: parent-to-child message, grid graph

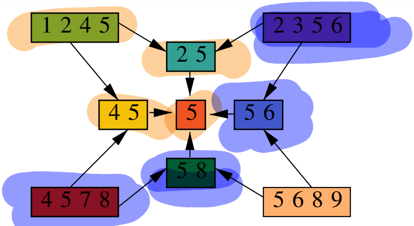




- Consider message for hyperedge $h = \{1, 2, 4, 5\}$, which has factors ψ' associated with (regular graph) edges $\{1, 2\}$, $\{2, 5\}$, $\{4, 5\}$, and $\{1, 4\}$ and also unary factors for each of the nodes 1, 2, 4, and 5 (eg., to associate evidence into the model).

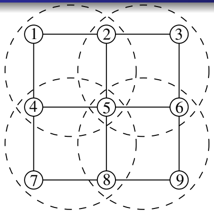
GBP examples: parent-to-child message, grid graph

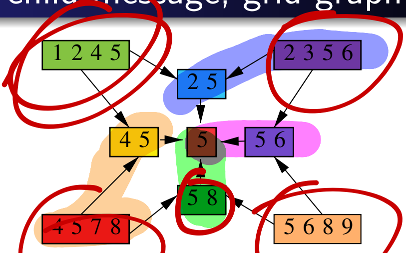




- Consider message for hyperedge $h = \{1, 2, 4, 5\}$, which has factors ψ' associated with (regular graph) edges $\{1, 2\}$, $\{2, 5\}$, $\{4, 5\}$, and $\{1, 4\}$ and also unary factors for each of the nodes 1, 2, 4, and 5 (eg., to associate evidence into the model).
- Then $\mathcal{D}^+(h) = \{\{1,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

GBP examples: parent-to-child message, grid graph

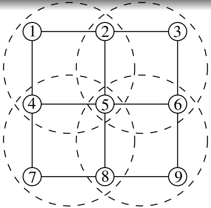


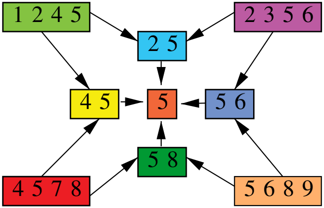


- Consider message for hyperedge $h = \{1,2,4,5\}$, which has factors ψ' associated with (regular graph) edges $\{1,2\}$, $\{2,5\}$, $\{4,5\}$, and $\{1,4\}$ and also unary factors for each of the nodes 1, 2, 4, and 5 (eg., to associate evidence into the model).
- Then $\mathcal{D}^+(h) = \{\{1,2,4,5\},\{4,5\},\{2,5\},\{5\}\}$.
- We get and expression for the marginal at h using the above formula.

$$\tau_{1,2,4,5} \propto \psi'_{1,2}\psi'_{1,4}\psi'_{2,5}\psi'_{4,5}\psi'_1\psi'_2\psi'_4\psi'_5 \times M_{\{2,3,5,6\}\to\{2,5\}} M_{\{4,5,7,8\}\to\{4,5\}} M_{\{5,6\}\to\{5\}} M_{\{5,8\}\to\{5\}} \quad (16.4)$$

GBP examples: parent-to-child message, grid graph





- Consider message for hyperedge $h = \{1, 2, 4, 5\}$, which has factors ψ' 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 $\mathcal{D}^+(h) = \{\{1, 2, 4, 5\}, \{4, 5\}, \{2, 5\}, \{5\}\}$.
- We get and expression for the marginal at h using the above formula.

$$\tau_{1,2,4,5} \propto \psi'_{1,2}\psi'_{1,4}\psi'_{2,5}\psi'_{4,5}\psi'_1\psi'_2\psi'_4\psi'_5 \qquad (16.4)$$
$$\times M_{\{2,3,5,6\}\to\{2,5\}} M_{\{4,5,7,8\}\to\{4,5\}} M_{\{5,6\}\to\{5\}} M_{\{5,8\}\to\{5\}}$$

- This could repeat for each of the largest clusters, until convergence.

Conjugate Duality, Maximum Likelihood, Negative Entropy

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:*

$$A^*(\mu) = \sup_{\theta \in \Omega} \left(\langle \theta, \mu \rangle - A(\theta) \right) = \begin{cases} -H(p_{\theta(\mu)}) & \text{if } \mu \in \mathcal{M}^\circ \\ +\infty & \text{if } \mu \notin \overline{\mathcal{M}} \end{cases} \tag{16.3}$$

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

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \left\{ \langle \theta, \mu \rangle - A^*(\mu) \right\} \tag{16.4}$$

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

$$\mu = \int_{\mathsf{D}_X} \phi(x) p_\theta(x) \nu(dx) = \mathbb{E}_\theta[\phi(X)] = \nabla A(\theta) \tag{16.5}$$

Expectation Propagation: basic idea

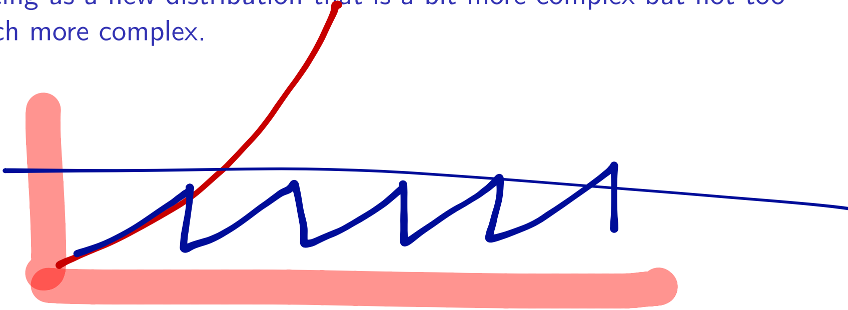
- Came from a method called "assumed density filtering" (ADF).

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.

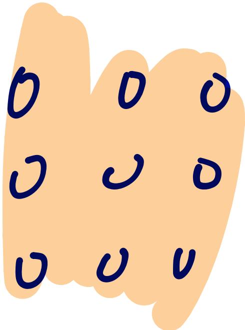
Expectation Propagation: basic idea

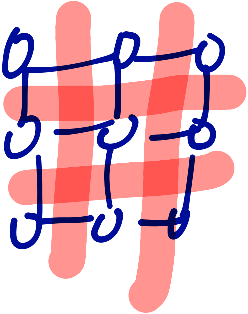
- Came from a method called "assumed density filtering" (ADF).
- Doing full inference involves exponential computation.
- We do a bit of inference, involving reasonable computation, and getting us a new distribution that is a bit more complex but not too much more complex.
- Before going further, we "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 **??**.

Term Decoupling

$\phi_\alpha(x)$ $\alpha \in \mathcal{I}$ $|\mathcal{I}| = d$

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





Term Decoupling

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

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

Term Decoupling

- Partition the d sufficient statistics into two parts, the tractable ones (of which there are d_T) and the intracxtable 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

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

Term Decoupling

- Partition the d sufficient statistics into two parts, the tractable ones (of which there are d_T) and the intracxtable 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

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

- ϕ_i are typically univariate, while Φ^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 intracxtable 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

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

- ϕ_i are typically univariate, while Φ^i are multivariate (b-dimensional).
- Consider exponential families associated with subcollection (ϕ, Φ).

Tractable component

- Tractable component

$$\phi \triangleq (\phi_1, \phi_2, \ldots, \phi_{d_T}) \tag{16.7}$$

Tractable component

- Tractable component

$$\phi \triangleq (\phi_1, \phi_2, \ldots, \phi_{d_T}) \tag{16.7}$$

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

Tractable component

- Tractable component

$$\phi \triangleq (\phi_1, \phi_2, \ldots, \phi_{d_T}) \tag{16.7}$$

- So $\phi : \mathcal{X}^m \to \mathbb{R}^{d_T}$ with vector of parameters $\theta \in \mathbb{R}^{d_T}$.
- Could instantiate model based only on this subcomponent, called the base model

Intractable component

- Intractable component

$$\Phi \triangleq (\Phi_1, \Phi_2, \ldots, \Phi_{d_I}) \tag{16.8}$$

Intractable component

- Intractable component

$$\Phi \triangleq (\Phi_1, \Phi_2, \ldots, \Phi_{d_I}) \tag{16.8}$$

- Each $\Phi_i : \mathcal{X}^m \to \mathbb{R}^b$.

Intractable component

- Intractable component

$$\Phi \triangleq (\Phi_1, \Phi_2, \ldots, \Phi_{d_I}) \tag{16.8}$$

- Each $\Phi_i : \mathcal{X}^m \to \mathbb{R}^b$.
- $\Phi : \mathcal{X}^m \to \mathbb{R}^{b \times d_I}$.

Intractable component

- Intractable component

$$\Phi \triangleq (\Phi_1, \Phi_2, \dots, \Phi_{d_I}) \tag{16.8}$$

- 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

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

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

Associated Distributions

- The associated exponential family

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

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

- Base model is tractable

$$p(x;\theta,\vec{0}) \propto \exp\left(\langle\theta,\phi(x)\rangle\right) \tag{16.11}$$

Associated Distributions

- The associated exponential family

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

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

- Base model is tractable

$$p(x;\theta,\vec{0}) \propto \exp\left(\langle \theta, \phi(x)\rangle\right) \tag{16.11}$$

- Φ^i-augmented model

$$p(x;\theta,\tilde{\theta}^i) \propto \exp\left(\langle \theta, \phi(x)\rangle\right) \exp\left(\left\langle \tilde{\theta}^i, \Phi^i(x)\right\rangle\right) \tag{16.12}$$

Associated Distributions: key points

The basic premises in the tractable-intractable partitioning between ϕ and Φ are:

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

Associated Distributions: key points

The basic premises in the tractable-intractable partitioning between ϕ and Φ are:

- It is possible to compute marginals exactly in polynomial time for distributions of the base form (any member of the ϕ-exponential family).
- For each $i = 1, \ldots, d_I$, exact polynomial-time computation is still possible for any Φ^i-augmented form (any member of the (ϕ, Φ^i)-exponential family).

Associated Distributions: key points

The basic premises in the tractable-intractable partitioning between ϕ and Φ are:

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

Example: Mixture models

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

Example: Mixture models

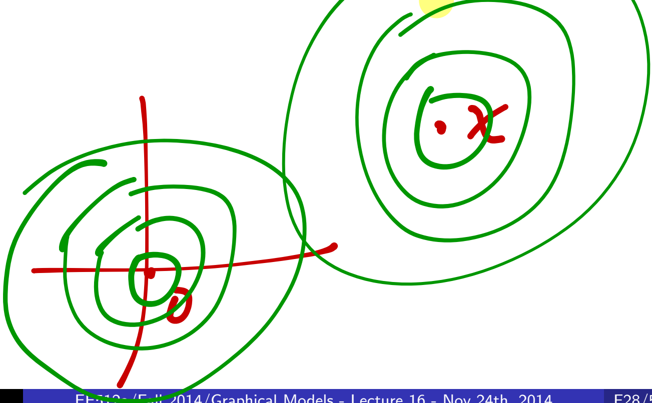
- Let $X \in \mathbb{R}^m$ be Gaussian with distribution $N(0, \Sigma)$.
- Let $\varphi(y; \mu, \Lambda)$ be Gaussian with mean μ covariance Λ.

x 0

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 μ covariance Λ.
- Suppose y conditioned on x is a two-component Gaussian mixture model taking the form:

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

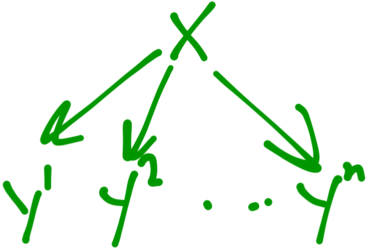


Example: Mixture models

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

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

- 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 μ covariance Λ.
- Suppose y conditioned on x is a two-component Gaussian mixture model taking the form:

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

- Assume we have obtained n i.i.d. samples $y^1, \dots, y^n$ from mixture density, and goal is to produce posterior $p(x|y^1, \dots, y^n)$, similar to Bayes-rule inverting a Naive-Bayes model.
- Using Bayes rule, we get mixture model with 2^n components!

$$p(x|y^1, \dots, y^n) \propto \exp\left(-\frac{1}{2}x^\intercal \Sigma^{-1} x\right) \prod_{i=1}^n p(y^i|X = x) \tag{16.14}$$

$$= \exp\left(-\frac{1}{2}x^\intercal \Sigma^{-1} x\right) \exp\left\{\sum_{i=1}^n \log p(y^i|X = x)\right\} \tag{16.15}$$

Example: Mixture models

- We equate $\exp\left(-\frac{1}{2}x^\intercal\sigma^{-1}x\right)$ with $\exp(\langle\theta,\phi(x)\rangle)$, with $d_T = m$.

Σ

Example: Mixture models

- We equate $\exp\left(-\frac{1}{2}x^\intercal \sigma^{-1} x\right)$ 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 $\exp\left(-\frac{1}{2}x^\intercal \sigma^{-1} x\right)$ 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 $\prod_{i=1}^{d_I} \exp\left(\left\langle \tilde{\theta}^i, \Phi^i(x)\right\rangle\right)$, with $b = 1$. These are the intractable factors.

Example: Mixture models

- We equate $\exp\left(-\frac{1}{2}x^\intercal\sigma^{-1}x\right)$ 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 $\prod_{i=1}^{d_I} \exp\left(\left\langle\tilde{\theta}^i, \Phi^i(x)\right\rangle\right)$, with $b = 1$. These are the intractable factors.
- Base distribution $p(x;\theta,\vec{0}) \propto \exp\left(-\frac{1}{2}x^\intercal\sigma^{-1}x\right)$ which is a Gaussian and easy as mentioned above.

Example: Mixture models

- We equate $\exp\left(-\frac{1}{2}x^\intercal\sigma^{-1}x\right)$ 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 $\prod_{i=1}^{d_I} \exp\left(\left\langle\tilde{\theta}^i, \Phi^i(x)\right\rangle\right)$, with $b = 1$. These are the intractable factors.
- Base distribution $p(x;\theta,\vec{0}) \propto \exp\left(-\frac{1}{2}x^\intercal\sigma^{-1}x\right)$ 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 $\exp\left(-\frac{1}{2}x^\intercal\sigma^{-1}x\right)$ 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 $\prod_{i=1}^{d_I} \exp\left(\left\langle\tilde{\theta}^i, \Phi^i(x)\right\rangle\right)$, with $b=1$. These are the intractable factors.
- Base distribution $p(x;\theta,\vec{0}) \propto \exp\left(-\frac{1}{2}x^\intercal\sigma^{-1}x\right)$ 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., Φ^i-augmented distribution is proportional to

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

Example: Mixture models

- We equate $\exp\left(-\frac{1}{2}x^\intercal\sigma^{-1}x\right)$ 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 $\prod_{i=1}^{d_I} \exp\left(\left\langle \tilde{\theta}^i, \Phi^i(x)\right\rangle\right)$, with $b = 1$. These are the intractable factors.
- Base distribution $p(x;\theta,\vec{0}) \propto \exp\left(-\frac{1}{2}x^\intercal\sigma^{-1}x\right)$ 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., Φ^i-augmented distribution is proportional to

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

- Computing marginals is easy (mixture of only 2 components)

Example: Mixture models

- We equate $\exp\left(-\frac{1}{2}x^\intercal \sigma^{-1} x\right)$ 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 $\prod_{i=1}^{d_I} \exp\left(\left\langle \tilde{\theta}^i, \Phi^i(x) \right\rangle\right)$, with $b = 1$. These are the intractable factors.
- Base distribution $p(x; \theta, \vec{0}) \propto \exp\left(-\frac{1}{2}x^\intercal \sigma^{-1} x\right)$ 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., Φ^i-augmented distribution is proportional to

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

- Computing marginals is easy (mixture of only 2 components)
- If we multiply in all Φ^i, becomes intractable (2^n potentially distinct components each of which requires marginalization).

Polytope and Base case

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

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

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

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

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

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

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

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

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

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

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

- 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 Φ^i-augmented exp. model and polytope

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

Augmented Base case

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

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

- 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 Φ^i-augmented exp. model and polytope

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

- 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).

Augmented Base case

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

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

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

$$\Pi^i(\tau, \tilde{\tau}) \to (\tau, \tilde{\tau}^i) \tag{16.20}$$

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"

$$\Pi^i(\tau, \tilde{\tau}) \to (\tau, \tilde{\tau}^i) \tag{16.20}$$

  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)

$$\mathcal{L}(\phi, \Phi) = \{(\tau, \tilde{\tau}) | \tau \in \mathcal{M}(\phi), \Pi^i(\tau, \tilde{\tau}) \in \mathcal{M}(\phi, \Phi^i), \ \forall i\} \tag{16.21}$$

New outer bound

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

$$\Pi^i(\tau, \tilde{\tau}) \rightarrow (\tau, \tilde{\tau}^i) \tag{16.20}$$

  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)

$$\mathcal{L}(\phi, \Phi) = \left\{(\tau, \tilde{\tau}) | \tau \in \mathcal{M}(\phi), \Pi^i(\tau, \tilde{\tau}) \in \mathcal{M}(\phi, \Phi^i), \ \forall i\right\} \tag{16.21}$$

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

New outer bound

- For any mean parms $(\tau, \tilde{\tau})$ where $\tilde{\tau} = (\tilde{\tau}^1, \tilde{\tau}^2, \ldots, \tilde{\tau}^{d_I})$, define coordinate "projection operation"
$$\Pi^i(\tau, \tilde{\tau}) \to (\tau, \tilde{\tau}^i) \tag{16.20}$$
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)
$$\mathcal{L}(\phi, \Phi) = \left\{(\tau, \tilde{\tau}) | \tau \in \mathcal{M}(\phi), \Pi^i(\tau, \tilde{\tau}) \in \mathcal{M}(\phi, \Phi^i),\ \forall i\right\} \tag{16.21}$$
- Note, based on a set of projections onto $\mathcal{M}(\phi, \Phi^i)$. Clearly outer bound since $\mathcal{M}(\phi, \Phi) \subseteq \mathcal{L}(\phi, \Phi)$.
- If Φ^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 ϕ-exponential family which mean parameters τ with entropy $H(\tau)$;

Members in new outer bound

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

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

Members in new outer bound

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

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

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

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

Members in new outer bound

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

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

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

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

- 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 ϕ-exponential family which mean parameters τ with entropy $H(\tau)$; B) Also, for $i = 1 \ldots d_I$, there is a member of the (ϕ, Φ^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:

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

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

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

- This characterizes the EP algorithms.
- Given graph $G = (V, E)$ when we take ϕ to be unaries V and Φ 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 $i \in [d_T]$.

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

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

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

$$\left\{\tau, (\eta^i, \tilde{\tau}^i), i \in [d_I]\right\} \in \mathbb{R}^{d_T} \times (\mathbb{R}^{d_T} \times \mathbb{R}^b)^{d_I} \tag{16.24}$$

- We arrive at the optimization:

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

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

$$\left\{\tau, (\eta^i, \tilde{\tau}^i), i \in [d_I]\right\} \in \mathbb{R}^{d_T} \times (\mathbb{R}^{d_T} \times \mathbb{R}^b)^{d_I} \tag{16.24}$$

- We arrive at the optimization:

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

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

$$L(\tau;\lambda) = \langle \tau, \theta \rangle + \sum_{i=1}^{d_I} \left\langle \tilde{\tau}^i, \tilde{\theta}^i \right\rangle + F(\tau; (\eta^i, \tilde{\tau}^i)) + \sum_{i=1}^{d_I} \langle \lambda^i, \tau - \eta^i \rangle + \dots \tag{16.26}$$

where

$$F(\tau; (\eta^i, \tilde{\tau}^i)) = H(\tau) + \sum_{i=1}^{d_I} \left[H(\eta^i, \tilde{\tau}^i) - H(\eta^i) \right] \tag{16.27}$$

and where λ^i are the Lagrange multipliers assocaited 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 $\left\{\tau, (\eta^i, \tilde{\tau}^i), i \in [d_I]\right\}$ 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. τ belongs to relative interior, i.e., $\tau \in \mathcal{M}^\circ(\theta)$ of the base model.

To Lagrangian optimization to Moment Matching

- Considering optimality conditions on what must hold for a solution $\left\{\tau, (\eta^i, \tilde{\tau}^i), i \in [d_I]\right\}$ to the above Lagrangian, must have properties:
  1. τ 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 $(\eta^i, \tilde{\tau}^i) \in \mathcal{M}^\circ(\phi, \Phi^i)$.

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

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

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

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

- Second condition means we're a member of the (ϕ, Φ^i)-exponential family, and (it can be shown) has form:

$$q^i(x, \theta, \tilde{\theta}^i, \lambda) \propto \exp\left(\left\langle \theta + \sum_{\ell \neq i} \lambda^\ell, \phi(x) \right\rangle + \left\langle \tilde{\theta}^i, \Phi^i(x) \right\rangle \right) \tag{16.29}$$

To Lagrangian optimization to Moment Matching

- Thid condiiton 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:

$$\int q(x;\theta,\lambda)\phi(x)\nu(dx) = \int q^i(x;\theta,\tilde{\theta}^i)\phi(x)\nu(dx) \tag{16.30}$$

fro $i \in [d_I]$.

Moment Matching → 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 $(\lambda^1, \dots, \lambda^{d_I})$
2. At each iteration $n = 1, 2, \dots$ choose some index $i(n) \in \{1, \dots, d_I\}$.

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

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

compute the mean parameters η^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 → Expectation Propagation Updates

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

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

compute the mean parameters η^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 $(\lambda^1, \dots, \lambda^{d_I})$
2. At each iteration $n = 1, 2, \dots$ choose some index $i(n) \in \{1, \dots, d_I\}$.
3. Under the following augmented distribution

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

compute the mean parameters η^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:

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

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:

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

- Then, each Φ^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

$$p(x; \theta, \theta_{u,v}) \propto (x; \theta, \vec{0}) \exp(\theta_{u,v}(x_u, x_v)) \tag{16.35}$$

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 Φ^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 Φ^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 Φ^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).

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 Φ^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 Φ^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. http://research.microsoft.com/en-us/um/people/minka/papers/

Mean Field

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

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.

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 $A(\theta)$.

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 $A(\theta)$.
- 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.

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 $A(\theta)$.
- 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.

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.

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.
- Ω gets smaller too. The parameters that respect F are of the form:

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

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.
- Ω gets smaller too. The parameters that respect F are of the form:

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

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 $\phi = (\phi_\alpha, \alpha \in \mathcal{I})$ 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.
- Ω gets smaller too. The parameters that respect F are of the form:

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

  notice, all parameters associated with sufficient statistic not in $\mathcal{I}(F)$ are set to zero, those statistics are nonexistent in F.
- If parameter was not zero, model would not respect the familiy 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

$$\Omega(F_0) = \left\{ \theta \in \Omega | \theta_{(s,t)} = 0 \;\; \forall (s,t) \in E(G) \right\} \tag{16.37}$$

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

$$\Omega(F_0) = \left\{\theta \in \Omega | \theta_{(s,t)} = 0 \ \ \forall (s,t) \in E(G)\right\} \tag{16.37}$$

- This is the all independence model, giving family of distributions

$$p_\theta(x) = \prod_{s \in V} p(x_s; \theta_s) \tag{16.38}$$

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 $\theta_{(s,t)}$.

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

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

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

$$\Omega(F_0) = \left\{\theta \in \Omega | \theta_{(s,t)} = 0 \;\; \forall (s,t) \notin T\right\} \tag{16.39}$$

- This gives a tree-dependent family

$$p_\theta(x) = \prod_{s \in V} p(x_s; \theta_s) \prod_{(s,t) \in T} \frac{p(x_s, x_t; \theta_{st})}{p(x_s; \theta_s) p(x_t; \theta_t)} \tag{16.40}$$

Inner bound Approximate Polytope

- 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 ϕ.

Inner bound Approximate Polytope

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

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

Inner bound Approximate Polytope

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

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

- 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.

Inner bound Approximate Polytope

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

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

- 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.
- Shorthand notation: $M_F^\circ(G) = M_F^\circ(G;\phi)$ 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:

$$A(\theta) \geq \langle \theta, \mu \rangle - A^*(\mu) \tag{16.43}$$

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

Mean field variational lower bound

Proof.

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

□

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:

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

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

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

□

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.

□

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 $A_{\text{MF}}(\theta)$ where:

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

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 μ is restricted in all cases, so $A_F^*(\mu)$ might be greatly simplified relative to $A^*(\mu)$.

Tractable Dual

- Normally dual $A^*(\mu) = \sup_{\theta \in \Omega} \left(\langle \theta, \mu \rangle - A(\theta) \right)$ 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 $A_{\mathsf{MF}}(\theta)$ where:

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

  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 μ 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.

Mean field, KL-Divergence, Exponential Model Families

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

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

Mean field, KL-Divergence, Exponential Model Families

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

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

- In summation form, we have

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

Mean field, KL-Divergence, Exponential Model Families

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

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

- In summation form, we have

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

- 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.

Mean field, KL-Divergence, Exponential Model Families

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

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

- In summation form, we have

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

- 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 θ or mean parameters μ. 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 $\theta^1, \theta^2 \in \Omega$

Mean field, KL-Divergence, Exponential Model Families

- 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 $\mu^i = \mathbb{E}_{\theta^i}[\phi(X)]$,

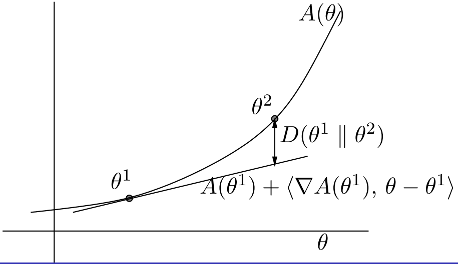
Mean field, KL-Divergence, Exponential Model Families

- 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:

$$D(\theta^1||\theta^2) = \mathbb{E}_{\theta^1}\left[\log \frac{p_{\theta^1}(x)}{p_{\theta^2}(x)}\right] \quad (16.50)$$

$$= A(\theta^2) - A(\theta^1) - \langle \mu^1, \theta^2 - \theta^1 \rangle \quad (16.51)$$

$$= A(\theta^2) - \left[A(\theta^1) + \langle \nabla A(\theta^1), \theta^2 - \theta^1 \rangle\right] \quad (16.52)$$



Mean field, KL-Divergence, Exponential Model Families

- Purely dual form of KL divergence can be formed as well, i.e.,

$$D(\theta^1||\theta^2) = D(\mu^1||\mu^2) = A^*(\mu^1) - A^*(\mu^2) - \langle \theta^2, \mu^1 - \mu^2 \rangle \quad (16.53)$$

Mean field, KL-Divergence, Exponential Model Families

- Purely dual form of KL divergence can be formed as well, i.e.,

$$D(\theta^1||\theta^2) = D(\mu^1||\mu^2) = A^*(\mu^1) - A^*(\mu^2) - \left\langle \theta^2, \mu^1 - \mu^2 \right\rangle \quad (16.53)$$

- Dual Bregman form

Mean field, KL-Divergence, Exponential Model Families

- Mixed/hybrid form of KL in terms of dual

Mean field, KL-Divergence, Exponential Model Families

- Mixed/hybrid form of KL in terms of dual
- We can also write the KL as:

$$D(\theta^1||\theta^2) = D(\mu^1||\theta^2) = A(\theta^2) + A^*(\mu^1) - \langle \mu^1, \theta^2 \rangle \tag{16.54}$$

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

Mean field, KL-Divergence, Exponential Model Families

- Mixed/hybrid form of KL in terms of dual
- We can also write the KL as:

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

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):

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

Mean field, KL-Divergence, Exponential Model Families

- Mixed/hybrid form of KL in terms of dual
- We can also write the KL as:

$$D(\theta^1||\theta^2) = D(\mu^1||\theta^2) = A(\theta^2) + A^*(\mu^1) - \langle \mu^1, \theta^2 \rangle \tag{16.54}$$

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):

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

- . . . can be written as:

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

Mean field, KL-Divergence, Exponential Model Families

- Since

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

Mean field, KL-Divergence, Exponential Model Families

- Since

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

- Thus, solving the mean-field variational problem of:

$$\max_{\mu \in \mathcal{M}_F(G)} \{\langle \mu, \theta \rangle - A_F^*(\mu)\} \tag{16.47}$$

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

Mean field, KL-Divergence, Exponential Model Families

- Since

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

- Thus, solving the mean-field variational problem of:

$$\max_{\mu \in \mathcal{M}_F(G)} \{\langle \mu, \theta \rangle - A_F^*(\mu)\} \tag{16.47}$$

  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_θ, over a family of "nice" distributions $M_F(G)$.

Naïve Mean field for Ising Model

- A classic example of mean-field (goes back to statistical physics)

Naïve Mean field for Ising Model

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

Naïve Mean field for Ising Model

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

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

Naïve Mean field for Ising Model

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

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

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

$$-A^*_{F_0}(\mu) = \sum_{s \in V} H_s(\mu_s) \tag{16.56}$$

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: $\mu_s = \mathbb{E}[X_s] = p(X_s = 1)$, $\mu_{st} = \mathbb{E}[X_s X_t] = p(X_s = 1, X_t = 1)$, thus $\mu \in \mathbb{R}^{|V|+|E|}$.
- Let $F_0 = (V, \emptyset)$ be our mean field approximation family. Thus,

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

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

$$-A^*_{F_0}(\mu) = \sum_{s \in V} H_s(\mu_s) \tag{16.56}$$

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

Naive Mean field for Ising Model

- We get variational lower bound problem

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

Naive Mean field for Ising Model

- We get variational lower bound problem

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

- 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

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

- Have constrained form of edge mean parameters $\mu_{st} = \mu_s \mu_t$
- $(\mu_1, \dots, \mu_m) \in [0,1]^m$ is m-D hypercube.

Naive Mean field for Ising Model

- We get variational lower bound problem

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

- 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.

Naive Mean field for Ising Model

- We get variational lower bound problem

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

- Have constrained form of edge mean parameters $\mu_{st} = \mu_s \mu_t$
- $(\mu_1, \dots, \mu_m) \in [0,1]^m$ 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

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

- 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 μ_s fixing all μ_t for $t \neq s$.

Naive Mean field for Ising Model

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

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

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 μ_s fixing all μ_t for $t \neq s$.
- Taking derivatives w.r.t. μ_s, we get the following update rule for element μ_s

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

  where $\sigma(z) = [1 + \exp(-z)]^{-1}$ is the sigmoid (logistic) function.
- This is the standard mean-field update that is quite well known, but derived from coordinate assent optimization of a variational perspective of the problem.

Naive Mean field for Ising Model

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

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

  where $\sigma(z) = [1 + \exp(-z)]^{-1}$ is the sigmoid (logistic) function.
- This is the standard mean-field update that is quite well known, but derived from coordinate assent 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

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

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

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

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

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 $q = p(X_1 = X_2)$.

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 $\mu_{12} = \mathbb{E}[X_1 X_2]$.
- Impose constraint $\mu_{12} = \mu_1 \mu_2$, we get mean field objective

$$f(\mu_1, \mu_2; \theta) = \theta_{12}\mu_1\mu_2 + \theta_1\mu_1 + \theta_2\mu_2 + H(\mu_1) + H(\mu_2) \quad (16.60)$$

  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:

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

  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

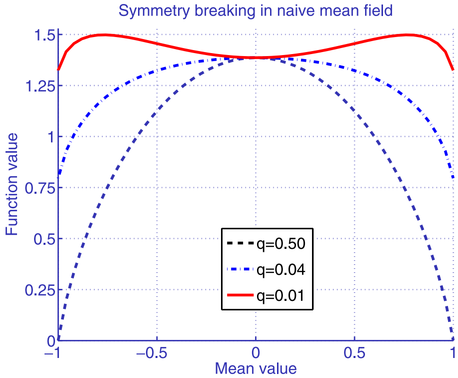
- For $q = 0.5$, objective $f(\mu_1, \mu_2; \theta(0.5))$ has global maximum at $(\mu_1, \mu_2) = (0, 0)$ so mean field is exact and convex. This corresponds to $p(X_1 = X_2) = 0$.

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

- Wainwright and Jordan *Graphical Models, Exponential Families, and Variational Inference* `http://www.nowpublishers.com/product.aspx?product=MAL&doi=2200000001`