Logistics

Announcements

- Should have read chapters 1, 2, 3, 4 in this book. Read chapter 5.
- Also should read “Divergence measures and message passing” by Thomas Minka, and “Structured Region Graphs: Morphing EP into GBP”, by Welling, Minka, and Teh.
- 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): Kikuchi, Expectation Propagation
- L17 (11/26): Expectation Propagation, Mean Field
- L18 (12/1):
- L19 (12/3):
- Final Presentations: (12/10):

Finals Week: Dec 8th-12th, 2014.

Drawing/Visualizing Hypergraphs as Bipartite Graphs

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

- It is possible to represent hypergraphs by only showing their hyperedges.
- Here, we see graphical representations of three hypergraphs. Subsets of nodes corresponding to hyperedges are shown in rectangles, whereas the arrows represent inclusion relations among hyperedges.

Which ones, if any, are in reduced representation?

Möbius Inversion Lemma and Inclusion-Exclusion

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

\begin{align*}
\forall A \subseteq V : \Upsilon(A) &= \sum_{B : B \subseteq A} \Omega(B) \quad (16.13) \\
\forall A \subseteq V : \Omega(A) &= \sum_{B : B \subseteq A} (-1)^{|A \setminus B|} \Upsilon(B) \quad (16.14)
\end{align*}

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 $P$ be a partially ordered set with binary relation $\leq$.
- A zeta function of a poset is a mapping $\zeta : P \times P \rightarrow \mathbb{R}$ defined by
  \[ \zeta(g, h) = \begin{cases} 
  1 & \text{if } g \leq h, \\
  0 & \text{otherwise.}
  \end{cases} \quad (16.23) \]
- The Möbius function $\omega : P \times P \rightarrow \mathbb{R}$ is the multiplicative inverse of this function. It is defined recursively:
  \[
  \begin{align*}
  \omega(g, g) &= 1 \quad \text{for all } g \in P \\
  \omega(g, h) &= 0 \quad \text{for all } h : h \not\leq g.
  \end{align*}
  \]
- Given $\omega(g, f)$ defined for $f$ such that $g \leq f < h$, we define
  \[ \omega(g, h) = -\sum_{\{f : g \leq f < h\}} \omega(g, f) \quad (16.24) \]
- Then, $\omega$ and $\zeta$ are multiplicative inverses, in that
  \[
  \sum_{f \in P} \omega(g, f)\zeta(f, h) = \sum_{\{f : g \leq f \leq h\}} \omega(g, f) = \delta(g, h) \quad (16.25)
  \]

General Möbius Inversion Lemma for Posets

**Lemma 16.2.8 (General Möbius Inversion Lemma)**

Given real valued functions $\Upsilon$ and $\Omega$ defined on poset $P$, then $\Omega(h)$ may be expressed via $\Upsilon(\cdot)$ via
\[
\Omega(h) = \sum_{g \leq h} \Upsilon(g) \quad \text{for all } h \in P \quad (16.23)
\]
iff $\Upsilon(h)$ may be expressed via $\Omega(\cdot)$ via
\[
\Upsilon(h) = \sum_{g \leq h} \Omega(g)\omega(g, h) \quad \text{for all } h \in P \quad (16.24)
\]

When $P = 2^V$ for some set $V$ (so this means that the poset consists of sets and all subsets of an underlying set $V$) this can be simplified, where $\leq$ becomes $\subseteq$; and $\geq$ becomes $\supseteq$, like we saw above.
(see Stanley, “Enumerative Combinatorics” for more info.)
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 \( \varphi_h \) 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.

Using Möbius, and Eqn. (16.23) 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 \( g \subseteq h \) 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 \( \mathbb{L}_t(G) \)
  \[
  \mathbb{L}_t(G) = \{ \tau \geq 0 \mid \text{Equations (16.3) } \forall h, \text{ and (16.34) } \forall g \subseteq h \text{ hold} \} \tag{16.35}
  \]

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:
  \[
  c(g) = \sum_{f \supseteq g} \omega(g, f) \tag{16.34}
  \]
### Variational Approach Amenable to Approximation

#### Variational Approximations we cover

- **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:

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

- 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.

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 \(H_{\text{app}} = \sum_{g \in \mathbb{E}} c(g) H_g(\tau_g)\) (via Möbius) to get Kikuchi variational approximation, message passing on hypergraphs.
3. Partition \(\tau\) into \((\tau, \tilde{\tau})\), and set \(\mathcal{M} \leftarrow \mathcal{L}(\phi, \Phi)\) and set \(-A^*(\mu) \leftarrow H_{\text{ep}}(\tau, \tilde{\tau})\) to get expectation propagation.

### Kikuchi variational approximation

- This at last gets the Kikuchi variational approximation

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

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

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

- We can define message passing algorithms on the hypertree, and show that if it converges, it is a fixed point of the associated Lagrangian.
Kikuchi variational approximation, 3x3 grid example

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

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

Kikuchi variational approximation, 3x3 grid example

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

- Complexity is only $O(r^4)$ and will stay $O(r^4)$ even as $n$ gets bigger (since clusters are at most size four).
Kikuchi and Hypertree-based Methods

**Generalized BP (GBP): Key idea**

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

\[
A_{Kikuchi}(\theta) = \max_{\tau \in \mathcal{L}(G)} \{\langle \theta, \tau \rangle + H_{\text{app}}(\tau)\} \tag{16.2}
\]

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

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

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

(16.3)

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

\[
\tau_{1,2,4,5} \propto \psi'_1 \psi'_2 \psi'_4 \psi'_5 \psi'_1 \psi'_2 \psi'_4 \psi'_5 \\
\times M_{\{2,3,5,6\} \to \{2,5\}} M_{\{4,5,7,8\} \to \{4,5\}} M_{\{5,6\} \to \{5\}} M_{\{5,8\} \to \{5\}}
\]

(16.4)

- 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:

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

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

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

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

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

---

**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 in EP

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

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

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

- $\phi_i$ are typically univariate, while $\Phi^i$ are typically multivariate ($b$-dimensional we’ll assume), although this need not always be the case (but will be for our exposition).
- Consider exponential families associated with subcollection $(\phi, \Phi)$.

Tractable component

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

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

- Intractable component

\[ \Phi \triangleq (\Phi_1, \Phi_2, \ldots, \Phi_d) \] (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: base and \( i \)-augmented

- The associated exponential family

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

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

- Base model is tractable

\[
p(x; \theta, \tilde{0}) \propto \exp (\langle \theta, \phi(x) \rangle) \tag{16.11}
\]

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

\[
p(x; \theta, \tilde{\theta}^i) \propto \exp (\langle \theta, \phi(x) \rangle) \exp \left( \langle \tilde{\theta}^i, \Phi^i(x) \rangle \right) \tag{16.12}
\]
Associated Distributions: key points

The basic premises in the tractable-intractable partitioning between $\phi$ and $\Phi$ are:

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

Example: Mixture models

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

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

(16.13)

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

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

(16.14)

\[
= \exp \left( -\frac{1}{2} x^T \Sigma^{-1} x \right) \exp \left\{ \sum_{i=1}^{n} \log p(y^i|X = x) \right\}
\]  

(16.15)
Example: Mixture models

- We equate $\exp\left(-\frac{1}{2}x^T\Sigma^{-1}x\right)$ with $\exp(\langle \theta, \phi(x) \rangle)$, with $d_T = m$.
- Such a distribution is multivariate Gaussian, and getting marginals (say $p(x_A)$ for $A \subseteq [m]$) from it is relatively “cheap” $O(m^3)$.
- $\exp \left\{ \sum_{i=1}^n \log p(y_i|X=x) \right\}$ equates to $\prod_{i=1}^{d_I} \exp \left( \langle \tilde{\theta}_i, \Phi^i(x) \rangle \right)$, with $b = 1$. These are the intractable factors.
- Base distribution $p(x; \theta, \vec{0}) \propto \exp \left(-\frac{1}{2}x^T\Sigma^{-1}x\right)$ which is a Gaussian and easy as mentioned above.
- If we multiply in only one intractable term, complexity to produce marginal still not so bad (quite easy in fact).
- I.e., $\Phi^i$-augmented distribution is proportional to
  \[
  \exp \left(-\frac{1}{2}x^T\Sigma^{-1}x\right) \left[(1 - \alpha)\varphi(y^i; 0, \sigma_0^2 I) + \alpha\varphi(y^i; x, \sigma_1^2 I)\right] \quad (16.16)
  \]
  Computing marginals is easy (mixture of only 2 components)
- If we multiply in all $\Phi^i$, becomes intractable ($2^n$ potentially distinct components each of which requires marginalization).

Polytope and Base case

- We can partition the mean parameters $(\mu, \tilde{\mu}) \in \mathbb{R}^{d_T + d_I \times b}$
- 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 $i = 1 \ldots d$ we have a $\Phi^i$-augmented exp. model and polytope

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

(16.19)

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

- Goal, variational approximation: Need outer bounds on $\mathcal{M}(\phi, \Phi)$ 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 EP-based outer bound

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

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

(16.20)

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

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

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

(16.21)

- Note, based on a set of projections onto $\mathcal{M}(\phi, \Phi^i)$.

- Outer bound, i.e., $\mathcal{M}(\phi, \Phi) \subseteq \mathcal{L}(\phi, \Phi)$, since:

$$
\tau \in \mathcal{M}(\phi) \iff \exists p \text{ s.t. } \tau = E_p[\phi(X)]
$$

(16.22)

$$
(\tau, \tilde{\tau}) \in \mathcal{L}(\phi, \Phi) \iff \tau \in \mathcal{M}(\phi) \& \exists p \text{ s.t. } (\tau, \tilde{\tau}^i) = E_p[\phi(X), \Phi^i(X)]
$$

(16.23)

$$
(\tau, \tilde{\tau}) \in \mathcal{M}(\phi, \Phi) \iff \exists p \text{ s.t. } (\tau, \tilde{\tau}) = E_p[\phi(X), \Phi(X)]
$$

(16.24)

- If $\Phi^i$ are edges of a graph (i.e. local consistency) then we get standard outer bound we saw before with Bethe approximation.
For any mean parms $(\tau, \hat{\tau}) \in \mathcal{L}(\phi, \Phi)$: A) There is a member of the $\phi$-exponential family which mean parameters $\tau$ with entropy $H(\tau)$; B) Also, for $i = 1 \ldots d_I$, there is a member of the $(\phi, \Phi^i)$-exponential family with mean parameters $(\tau, \hat{\tau}^i)$ with entropy $H(\tau, \hat{\tau}^i)$.

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

$$H(\tau, \hat{\tau}) \approx H_{ep}(\tau, \hat{\tau}) \overset{\Delta}{=} H(\tau) + \sum_{\ell=1}^{d_I} \left[ H(\tau, \hat{\tau}^l) - H(\tau) \right]$$  \hspace{1cm} (16.25)

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

$$\max_{(\tau, \hat{\tau}) \in \mathcal{L}(\phi, \Phi)} \left\{ \langle \tau, \theta \rangle + \langle \hat{\tau}, \hat{\theta} \rangle + H_{ep}(\tau, \hat{\tau}) \right\}$$  \hspace{1cm} (16.26)

This characterizes the EP algorithms.

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

Lagrangian optimization setup

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

This gives large set of pseudo-mean parameters

$$\{ \tau, (\eta^i, \hat{\tau}^i), i \in [d_I] \} \in \mathbb{R}^{dT} \times \left( \mathbb{R}^{dT} \times \mathbb{R}^b \right)^{d_I}$$  \hspace{1cm} (16.27)

We arrive at the optimization:

$$\max_{\{ \tau, \{ (\eta^i, \hat{\tau}^i) \} \}_i} \left\{ \langle \tau, \theta \rangle + \sum_{i=1}^{d_I} \langle \hat{\tau}^i, \hat{\theta}^i \rangle + H(\tau) + \sum_{i=1}^{d_I} \left[ H(\eta^i, \hat{\tau}^i) - H(\eta^i) \right] \right\}$$  \hspace{1cm} (16.28)

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

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

- We get a Lagrangian version of the objective

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

(16.29)

where

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

(16.30)

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

To Lagrangian optimization to Moment Matching

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

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

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

(16.31)

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

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

(16.32)
To Lagrangian optimization to Moment Matching

- This condition is a form of moment-matching. I.e., we have $\tau = E_{q}[\phi(X)]$ and $\eta^i = E_{q^i}[\phi(X)]$, so equating these gives:

$$\int q(x; \theta, \lambda)\phi(x)\nu(dx) = \int q^i(x; \theta, \tilde{\theta}^i)\phi(x)\nu(dx) \quad (16.33)$$

for $i \in [d_I]$.

Moment Matching $\rightarrow$ Expectation Propagation Updates

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

$$q^i(x; \theta, \tilde{\theta}^i, \lambda) \propto \exp \left( \left[ \theta + \sum_{\ell \neq i} \lambda^\ell \phi(x) \right] + \left[ \tilde{\theta}^i, \Phi^i(x) \right] \right), \quad (16.34)$$

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

$$\eta^{i(n)} = \int q^{i(n)}(x)\phi(x)\nu(dx) = E_{q^{i(n)}}[\phi(X)] \quad (16.35)$$

4. Form base distribution $q$ using Equation 16.31 and adjust $\lambda^{i(n)}$ to satisfy the moment-matching condition

$$E_{q}[\phi(X)] = \eta^{i(n)} \quad (16.36)$$

5. This is a KL-divergence minimization step, but done w. exponential family models which thus corresponds to moment-matching.
Sources for Today’s Lecture