Outstanding Reading

- Continue chapter 7 in the Wainwright and Jordan book.
- Start chapter 8 in the Wainwright and Jordan book.
We need to find one makeup lecture this term.

- L1 (9/28): Introduction, Families, Semantics
- LX (9/30): No class
- L2 (10/5): Trees, exact inference
- L3 (10/7): More on trees and inference.
- L4 (10/12): To tree or not to tree.
- L5 (10/14): All models lead to trees
- L6 (10/19): Decomposable, JT
- L7 (10/21): Inference on JT’s
- L8 (10/26): JT Inference, semi-rings,
- L9 (10/28): time-space tradeoff, conditioning, LBP
- L10 (11/2): LBP, exp. f. models
- L11 (11/4): exp. f. models, marg poly
- L12 (11/9): pseudo marg, Bethe
- LXX (11/11): Veterans Day, no class
- L13 (11/16): Bethe, loop series
- LXX (11/25): Thanksgiving, no class
- L16 (11/30): mean field
- L17 (12/2): convexified, tree reweighted
- L18 (12/7): tree reweighted, MPE
- L19: (12/9): Final Presentations: (12/9):

---

**Theorem 1 (Relationship between $A$ and $A^*$)**

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

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

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

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

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

$$\mu = \int_{D_X} \phi(x)p_{\theta}(x)\nu(dx) = \mathbb{E}_{\theta}[\phi(X)] = \nabla A(\theta) \quad (3)$$
**Variational Problem**

Original variational representation of log partition function

\[
A(\theta) = \sup_{\mu \in \mathcal{M}} \{ \langle \theta, \mu \rangle - A^*(\mu) \} \tag{4}
\]

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

\[
A(\theta) \geq \max_{\mu \in \mathcal{M}_F(G)} \{ \langle \mu, \theta \rangle - A^*_F(\mu) \} = A_{\text{mf}}(\theta) \tag{5}
\]

since \( \mathcal{M}_F(G) \subseteq \mathcal{M} \)

---

**Convex Relaxations and Upper Bounds**

\[
A(\theta) = \sup_{\mu \in \mathcal{M}} \{ \langle \theta, \mu \rangle - A^*(\mu) \} \tag{6}
\]

- What about upper bounds?
- Other than mean field, none of the other approximation methods have been anything other than approximation methods.
- We would like both lower and upper bounds of \( A(\theta) \) since that will allow us to produce upper and lower bounds of the probabilistic queries we wish to perform.
- If the upper and lower bounds between a given probably \( p \) is small, \( p_L \leq p \leq p_U \), with \( p_U - p_L \leq \epsilon \), we have guarantees, for a particular instance of a model.
- In this next chapter (Chap 7), we will “convexify” \( H(\mu) \) and at the same time produce upper bounds.
Recall sufficient stats $\phi = (\phi_\alpha, \alpha \in I)$ and canonical parameters $\theta = (\theta_\alpha, \alpha \in I)$.

In general, inference (computing mean parameters) is hard for a given $G$.

For a tractable subgraph $F$, it is not so hard, as we saw in the mean field case. Note in mean field case, we had one particular $F$.

Let $D$ be a set of subfamilies that are tractable.

I.e., $D$ might be all spanning trees of $G$, or some subset of spanning trees that we like.

As before, $I(F) \subseteq I$ are the indices of the suff. stats. that abide by $F$, and $|I(F)| = d(F) < d = |I|$ suff. stats.

As before, $\mathcal{M}(F)$ is set of realizable mean parameters associated with $F$, so that $\mu(F) \in \mathcal{M}(F)$. Thus, $\mathcal{M}(F) \subseteq \mathbb{R}^{|I(F)|}$, and

$$\mathcal{M}(F) = \left\{ \mu \in \mathbb{R}^{|I(F)|} | \exists \mu_s.t. \mu_\alpha = \mathbb{E}_p[\phi_\alpha(X)] \ \forall \alpha \in I(F) \right\} \quad (7)$$

Note $\mathcal{M}_F(G) \neq \mathcal{M}(F)$.

Given $\mu \in \mathcal{M}$, $\mu(F) \in \mathcal{M}(F)$ projects from $I$ to $I(F)$.

Thus, for any $\mu \in \mathcal{M} \subseteq \mathbb{R}^d$, we have that $\mu(F) \in \mathcal{M}(F) \subseteq \mathbb{R}^{d(F)}$.

We can moreover define the entropy associated with projected mean, namely $H(\mu(F)) \triangleq H(p_{\mu(F)}) = -A^*(\mu(F))$.

Critically, we have that $H(\mu(F)) \geq H(\mu) = H(p_{\mu})$, as we show next.
Proposition 2

**Maximum Entropy Bounds** Given any mean parameter \( \mu \in \mathcal{M} \) and its projection \( \mu(F) \) onto any subgraph \( F \), we have the bound

\[
A^*(\mu(F)) \leq A^*(\mu)
\]  

or alternatively stated, \( H(\mu(F)) \geq H(\mu) \).

- Intuition: \( H(\mu) = H(p_\mu) \) is the entropy of the exponential family model with mean parameters \( \mu \).
- equivalently \( H(\mu) = H(p_\mu) \) is the entropy of the distribution that is the solution to the maximum entropy problem subject to the constraints that it has \( \mu = \mathbb{E}_{p_\theta}[\phi(X)] \).
- When we form \( \mu(F) \), there are fewer constraints, so the entropy in the corresponding maximum entropy problem may get larger.
- Thus, \( H(\mu(F)) \geq H(\mu) \).

**Proof.**

- Dual problem

\[
A^*(\mu) = \sup_{\theta \in \mathbb{R}^d} \{ \langle \mu, \theta \rangle - A(\theta) \}
\]  

- Dual problem in sub-graph case.

\[
A^*(\mu(F)) = \sup_{\theta(F) \in \mathbb{R}^d(F)} \{ \langle \mu(F), \theta(F) \rangle - A(\theta(F)) \}
\]  

- Dual problem in sub-graph case — alternate expression

\[
A^*(\mu(F)) = \sup_{\theta \in \mathbb{R}^d} \{ \langle \mu, \theta \rangle - A(\theta) \}
\]

\[
\theta_\alpha = 0 \ \forall \alpha \notin \mathcal{I}(F)
\]

Thus, \( A^*(\mu) \geq A^*(\mu(F)) \).
Convex Relaxations and Upper Bounds - Relaxed Entropy

- Note that the upper bound is true for each $F \in \mathcal{D}$, and thus would be true for mixtures of different $F \in \mathcal{D}$.

- We can form a distribution $\rho$ over tractable structures. I.e., $\rho \in \mathbb{R}^{\lvert \mathcal{D} \rvert}$, i.e., $\rho(F) \geq 0$ for $F \in \mathcal{D}$ and $\sum_{F \in \mathcal{D}} \rho(F) = 1$.

- Convex combination, gives general upper bound

$$H(\mu) \leq \mathbb{E}_\rho[H(\mu(F))] = \sum_{F \in \mathcal{D}} \rho(F)H(\mu(F)) \quad (12)$$

- This will be our convexified upper bound on entropy.

- Compared to mean field, we are not choosing only one structure, but many of them, and mixing them together in certain ways.

Convex Relaxations and Upper Bounds - Outer bound

- When we form the mixture, and we wish to evaluate a given $\mu(F)$ on it, we need to make sure that each component can properly evaluate any possible $\mu(F)$, so logical constraint is to make sure any $\mu(F)$ works for all of them.

- Constraint set as follows:

$$\mathcal{L}(G; \mathcal{D}) = \left\{ \tau \in \mathbb{R}^d \mid \tau(F) \in \mathcal{M}(F) \quad \forall F \in \mathcal{D} \right\} \quad (13)$$

$$= \bigcap_{F \in \mathcal{D}} \mathcal{M}(F) \quad (14)$$

- Note this is an outer bound i.e., $\mathcal{L}(G; \mathcal{D}) \supseteq \mathcal{M}(G)$ since any member of $\mathcal{M}(G)$ (any valid mean parameter for $G$) must also be a member of any $\mathcal{M}(F)$ (i.e., non-neg, sums to 1, and consistency).

- Also note, $\mathcal{L}(G; \mathcal{D})$ is convex since it is the intersection of a set of convex sets.
Convex Upper Bounds

- Combining the upper bound on entropy, and the outer bound on $M$, we get a new variational approximation to the cumulant function.

\[ B_D(\theta; \rho) \triangleq \sup_{\tau \in \mathcal{L}(G; \mathcal{D})} \left\{ \langle \tau, \theta \rangle + \sum_{F \in \mathcal{D}} \rho(F) H(\tau(F)) \right\} \quad (15) \]

- Objective is convex in $\theta$ since it is a max over a set of affine functions of $\theta$ (i.e., $g(\theta) = \max_{\tau} \langle \tau, \theta \rangle + c_\tau$)
- Also, $\mathcal{L}(G; \mathcal{D})$ is a convex outer bound on $M(G)$
- Thus $B_D(\theta; \rho)$ is convex, has a global optimal solution, it approximates $A(\theta)$, and best of all is an upper bound, $A(\theta) \leq B_D(\theta; \rho)$

Tree-reweighted sum-product and Bethe

- We can get convex upper bounds in the tree case, and a new style of sum-product algorithm.
- Consider MRF again

\[ p_\theta(x) \propto \exp \left\{ \sum_{s \in V} \theta_s(x_s) + \sum_{(s,t) \in E} \theta_{st}(x_s, x_t) \right\} \quad (16) \]

- Let $\mathcal{T}$ be a set of all spanning trees $T$ of $G$, and let $\rho$ be a distribution over them, $\sum_{T \in \mathcal{T}} \rho(T) = 1$.
- Thus, we have $H(\mu) \leq \sum_{T \in \mathcal{T}} \rho(T) H(\mu(T))$
- For any $T$, $H(\mu(T))$ has an easy form, i.e.,

\[ H(\mu(T)) = \sum_{s \in V} H_s(\mu_s) - \sum_{(s,t) \in E(T)} I_{st}(\mu_{st}) \quad (17) \]

- We want to use this to see what happens when we take the expected value w.r.t. distribution $\rho$. 
Every tree is spanning, all trees have all node, so the probability, according to $\rho$ if a given node is always 1. i.e., $\rho_s = 1, \forall s \in V$.

Thus, in $\mathbb{E}_{\rho}[H(\mu(T))]$, we have a term of the form $\sum_{s \in V} H_s(\mu_s)$.

For edges we need $\rho_{st} = \mathbb{E}_{\rho}[\mathbb{I}((s, t) \in E(T))]$, this indicates the probability of presence of an edge in the set $\mathcal{E}$.

The expression becomes

$$H(\mu) \leq \sum_{s \in V} H_s(\mu_s) - \sum_{(s, t) \in E} \rho_{st} I_{st}(\mu_{st}) \quad (18)$$

Note right hand sum is over all $E$ (not just a given spanning tree) and terms are weighted by probability of the given edge $\rho_{st}$.

We also need outer bound on $\mathcal{M}$.

For discrete case $\mathcal{M} = \mathcal{M}(G)$ is marginal polytope.

$\mathcal{M}(T)$ is marginal polytope for tree, and for a tree is the same as $\mathcal{L}(T)$, the locally consistent pseudo-marginals (which recall are marginals for a tree).

Thus, $\mu(T) \in \mathcal{M}(T)$ requires non-negativity, sum-to-one (at each node), and edge-to-node consistency (marginalization) on each edge. If $G = T$ then we’re done.

For general $G$, If we ask for $\mu(T) \in \mathcal{M}(T)$ for all $T \in \mathcal{I}$, this is identical to asking for local marginalization on every edge of $G$.

Thus, in this case $\mathcal{L}(G; \mathcal{I})$ is just the set of locally consistent pseudomarginals, and is the same as the outer bound we saw in the Bethe variational approximation $\mathcal{L}(G)$.

In Bethe case, however, we did not have a bound on entropy, only an outer bound on the marginal polytope. Now, however, we also have a (convexification based) bound on entropy.
Theorem 3 (Tree-Reweighted Bethe and Sum-Product)

(a) For any choice of edge appearance vector \((\rho_{st}, (s, t) \in E)\) in the spanning tree polytope, the cumulant function \(A(\theta)\) evaluated at \(\theta\) is upper bounded by the solution of the tree reweighted Bethe variational problem (BVP):

\[
B_T(\theta; \rho) = \max_{\tau \in \mathbb{L}(G)} \left\{ \langle \tau, \theta \rangle + \sum_{s \in V} H_s(\tau_s) - \sum_{(s,t) \in E} \rho_{st} I_{st}(\tau_{st}) \right\}
\]

For any edge appearance vector such that \(\rho_{st} > 0\) for all edges \((s, t)\), this problem is strictly convex with a unique optimum.

(b) The tree-reweighted BVP can be solved using the tree-reweighted sum-product updates

\[
M_{t \to s}(x_s) \leftarrow \kappa \sum_{x'_t \in \mathcal{X}_t} \varphi_{st}(x_s, x'_t) \prod_{v \in N(t) \setminus s} \frac{[M_{v \to t}(x'_t)]^{\rho_{vt}}}{[M_{s \to t}(x'_t)]^{1-\rho_{ts}}} \]

where \(\varphi_{st}(x_s, x'_t) = \exp\left(\frac{1}{\rho_{st}} \phi_{st}(x_s, x'_t) + \theta_t(x'_t)\right)\). The updates have a unique fixed point under (a).
Tree-reweighted sum-product and Bethe

- Note that if $\rho_{st} \leftarrow 1$, for all $(s, t) \in E$, then we recover standard LBP and Bethe approximation.
- Thus, this is a true convex generalization.
- However, if $\rho_{st} = 1$ then edge $(s, t)$ appears in all spanning trees. If this is indeed true for all spanning trees $T$, then $G$ must be a tree, and we get back standard tree-based message passing we saw in lecture 2!!

Tree-reweighted sum-product example

(a) a graph, and (b), (c), and (d) various spanning trees.

- What are the edge probabilities $\rho_{st}$?
In above case, we have both a convexification of the cumulant and an upper bound property.

It should be pointed out that these are not mutual requirements: one can have convex without upper bound and vice versa.

The fixed point has following form:

$$\tau^*_s(x_s) = \kappa \exp \left\{ \theta_s(x_s) \right\} \prod_{v \in N(s)} [M^*_{v \to s}(x_s)]^{\rho_{vs}}$$  \hspace{1cm} (21)

$$\tau^*_{st}(x_s, x_t) = \kappa \varphi_{st}(x_s, x_t) \frac{\prod_{v \in N(s) \setminus t} [M^*_{vs}(x_s)]^{\rho_{vs}} \prod_{v \in N(t) \setminus s} [M^*_{vt}(x_t)]^{\rho_{vt}}}{[M^*_{ts}(x_s)](1-\rho_{st}) [M^*_{st}(x_t)](1-\rho_{ts})}$$  \hspace{1cm} (22)

with \( \varphi_{st}(x_s, x_t) = \exp \left\{ \frac{1}{\rho_{st}} \theta_{st}(x_s, x_t) + \theta_s(x_s) + \theta_t(x_t) \right\} \)

Damping appears in practice to help convergence, where each new message is a convex mixture of the previous version if itself and the new message according to the equations.
Why stop at trees, instead could use hypertrees and then deduce a hypertree version of the reweighted BP algorithm.

Example in book considers $k$-trees, with tree width at most $t$. I.e. $\Xi(t)$.

Then we get the same form of bounds

\[ H(\mu) \leq E\rho[H(\mu(T))] = - \sum_{T \in \Xi(t)} \rho(T) H(\mu(T)) \]  

but here $T$ is over all valid $k$-trees.

This leads to a convexified Kikuchi variational problem

\[ A(\theta) \leq B_{\Xi(t)}(\theta; \rho) = \max_{\tau \in L(G)} \{ \langle \tau, \theta \rangle + E\rho[H(\tau(T))] \} \]  

same form as before, but of course quite different.

Other variational variants have convexified version.

Convexified forms of EP

\[ H_{ep}(\tau, \tilde{\tau}; \rho) = H(\tau) + \sum_{\ell=1}^{d_1} \rho(\ell) [H(\tau, \tilde{\tau}^\ell) - H(\tau)] \]  

where $\sum_\ell \rho(\ell) = 1$.

Lagrangian formulation, solutions to that, reweighted EP, "power EP"
Other variants

- Why only trees? There could be other tractable families.
- Planar graphs, restricted grids
- Other forms, perhaps it would be possible to take mixtures of structures each of which might not have low tree width but has restricted potentials in some way.
- Examples from book:
  - Use of Gaussian continuous entropy as an upper bound and a covariance-based outer bound of $M$.
  - Use of conditional entropy, various forms of use of polyhedral approximations.
- This is a very active research area right now.

MPE - most probable explanation

- In many cases, we care not to sum over $x$ in $\sum_x p(x)$ but instead to compute $x^* \in \text{argmax}_{x \in D_X} p(x)$.
- This is called the “Viterbi assignment”, or the “most probable explanation” (MPE), or the “most probable configuration” or the “mode”, or a few other names.
- From the perspective of semirings, we are only changing the semiring (from sum-product to max-product). Can do exactly same form of exact inference algorithms (e.g., trees, $k$-trees, junction trees) using different semiring, to get answer. To get $n$-best answers, can also be seen as a semiring.
- Equally difficult when tree-width is large.
- Can the variational approach help in this case as well?
MPE - most probable explanation

- MPE again
\[
\arg\max_{x \in D_X^m} p(x) = \{x \in D_X^m : p_{\theta}(x) \geq p_{\theta}(y), \forall y \in D_X^m\} \tag{26}
\]

- Since we are using exponential family models, we have
\[
\arg\max_{x \in D_X^m} p(x) = \arg\max_{x \in D_X^m} \langle \theta, \phi(x) \rangle \tag{27}
\]
\[
\text{i.e., cumulant function isn’t required for computation.}
\]
- But it is related. Recall cumulant function
\[
A(\theta) = \log \int \exp \{\langle \theta, \phi(x) \rangle\} d\nu(x) \tag{28}
\]
\[
= \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{29}
\]

MPE - and variational

- Considering \( p_{\theta}(x) = \exp \{\langle \theta, \phi(x) \rangle - A(\theta)\} \).
- Let \( \beta \in \mathbb{R}_+ \) be a positive scalar.
- If we substitute \( \theta \) with \( \beta \theta \) (i.e., \( p_{\theta}(x) \) with \( p_{\beta \theta}(x) \)), then \( b_{\beta \theta}(x) \) becomes more concentrated around MPE solutions as \( \beta \to \infty \).
- This should have some influence on the cumulant. I.e., if we look at \( A(\beta \theta) / \beta \) and let \( \beta \) get large.
- Moreover, with respect to mean parameters, the maximum mean should (intuitively) fall on a vertex.
We have following theorem.

**Theorem 4**

*For all* $\theta \in \Omega$, *the problem of mode computation has the following alternative representations:*

$$\max_{x \in D_{X^m}} \langle \theta, \phi(x) \rangle = \max_{\mu \in \mathcal{M}} \langle \theta, \mu \rangle, \text{ and}$$

$$\max_{x \in D_{X^m}} \langle \theta, \phi(x) \rangle = \lim_{\beta \to \infty} \frac{A(\beta \theta)}{\beta} \quad (31)$$

- First equation shows how MPE can be seen as a LP over convex set $\mathcal{M}$.
- For discrete distributions, we have $\mathcal{M}(G)$ for graph $G$, so this is a linear objective with polyhedral constraints.
- Since l.h.s. is IP, this shows the difficulty of $\mathcal{M}(G)$.

---

**MPE - and variational for trees**

- When the graph is a tree, we can find an interesting connection between the max-product form of messages and a particular Lagrangian.
- Using the above theorem, we get (for tree $T$)

$$\max_{x \in D_{X^m}} \left[ \sum_{s \in V} \theta_s(x_s) + \sum_{(s,t) \in E} \theta_{st}(x_s, x_t) \right] = \max_{\mu \in \mathcal{L}(T)} \langle \mu, \theta \rangle \quad (32)$$

- Right hand side is a LP over a simple polytope, the marginal polytope for trees $\mathcal{L}(T)$.
- It turns out that: the max-product updates are a Lagrangian method for solving dual of the above linear program.
- Maxproduct updates take the form:

$$M_{t \to s}(x_s) \leftarrow \kappa \max_{x_t \in D_{X_t}} \left[ \exp \left\{ \theta_{st}(x_s, x_t) + \theta_t(x_t) \right\} \prod_{u \in N(t) \setminus s} M_{u \to t}(x_t) \right] \quad (33)$$
Max-Product and LP Duality

**Theorem 5**

Max-product and LP Duality Consider the dual function $Q$ defined by the following partial Lagrangian formulation of the tree-structured LP:

$$Q(\lambda) = \max_{\mu \in \mathbb{N}} L(\mu; \lambda), \quad \text{where}$$

$$L(\mu; \lambda) = \langle \theta, \mu \rangle + \sum_{(s,t) \in E(T)} \left[ \sum_{x_s} \lambda_{ts}(x_s)C_{ts}(x_s) + \sum_{x_t} \lambda_{st}(x_t)C_{st}(x_t) \right]$$

(35)

For any fixed point $M^*$ of the max-product updates, the vector $\lambda^* = \log M^*$, where the logarithm is taken elementwise, is an optimal solution of the dual problem $\min_{\lambda} Q(\lambda)$. 

---

**Scratch Paper**
Sources for Today’s Lecture

- this material comes from the Wainwright and Jordan book.