Outstanding Reading

- Finish chapter 5 in the Wainwright and Jordan book.
- Read chapter 7 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 JTs
- L8 (10/26): JT Inference, semi-rings,
- L9 (10/28): time-space tradeoff, conditioning, LBP
- L10 (11/2): LBP, exp. f. models
  - L11 (11/4): exp. f. models, marg poly
  - L12 (11/9): pseudo marg, Bethe
  - LXX (11/11): Veterans Day, no class
  - L13 (11/16): Bethe, loop series
  - LXX (11/25): Thanksgiving, no class
  - L16 (11/30): mean field
  - L17 (12/2): convexified, tree reweighted
  - L18 (12/7):
  - L19: (12/9):
  - Final Presentations: (12/9):

**Conjugate Duality**

**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:

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

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

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

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

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

Original variational representation of log partition function

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

- Set \( M \leftarrow \mathbb{L} \) and \(-A^*(\mu) \leftarrow H_{\text{Bethe}}(\tau)\) to get Bethe variational approximation, LBP fixed point.
- Set \( M \leftarrow \mathbb{L}_t(G) \) (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 \( M \leftarrow \mathcal{L}(\phi, \Phi) \) and set \(-A^*(\mu) \leftarrow H_{\text{ep}}(\tau, \tilde{\tau})\) to get expectation propagation.
- Mean field (from variational approximation perspective) is

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

Mean Field

- So far, we have been using an outer bound on \( M \).
- In mean-field methods, we use an “inner bound”, a subset of \( M \) 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 subgraph $F$ that is a spanning subgraph.
- Simplest example: $F = (V, \emptyset)$ all independence model.
- Tree example: $F = (V, E_T)$ where edges $E_T \subset E$ constitute a spanning tree.
- Exponential family, sufficient statistics $\phi = (\phi_\alpha, \alpha \in I)$ associated with this family $I(F) \subseteq I$. These are the statistics that respect the Markov properties of subgraph $F$.
- $\Omega$ gets smaller too. The parameters that respect $F$ are of the form:

$$\Omega(F) \triangleq \{ \theta \in \Omega | \theta_\alpha = 0 \ \forall \alpha \in I \setminus I(F) \}$$  \hspace{1cm} (6)

notice, all parameters associated with sufficient statistic not in $I(F)$ are set to zero, so those statistics are essentially “turned off.”

Inner bound Approximate Polytope

- Before, we had $\mathcal{M}(G; \phi)$, all possible mean parameters associated with $G$ and associated set of sufficient statistics $\phi$.
- For a given subgraph $F$, we only consider those mean parameters possible under such models. I.e.,

$$\mathcal{M}_F(G; \phi) = \left\{ \mu \in \mathbb{R}^d | \mu = \mathbb{E}_\theta[\phi(x)] \text{ for some } \theta \in \Omega(F) \right\}$$  \hspace{1cm} (7)

- Therefore, we have that

$$\mathcal{M}^o_F(G; \phi) \subseteq \mathcal{M}^o(G; \phi)$$  \hspace{1cm} (8)

and so $\mathcal{M}^o_F(G; \phi)$ is an inner approximation of the set of realizable mean parameters.
- Shorthand notation: $M^o_F(G) = M^o_F(G; \phi)$ and $M^o(G) = M^o(G; \phi)$

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 2 (mean field lower bound)**

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

$$A(\theta) \geq \langle \theta, \mu \rangle - A^*(\mu) \quad (9)$$

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

Tractable Dual

- Normally dual $A^*$ is intractable, 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 do

$$A(\theta) \geq \max_{\mu \in \mathcal{M}_F(G)} \{ \langle \mu, \theta \rangle - A^*_F(\mu) \} = A_{mf}(\theta) \quad (10)$$

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

- Note, for $\mu \in \mathcal{M}_F(G)$, $A^*_F(\mu)$ is not an approximation.
Mean field, KL-Divergence, Exponential Model Families

- We can also write the KL as:

\[ D(\theta^1 || \theta^2) = D(\mu^1 || \theta^2) = A(\theta^2) + A^*(\mu^1) - \langle \mu^1, \theta^2 \rangle \quad (11) \]

via \( A^*(\mu^1) = \langle \theta^1, \mu^1 \rangle - A(\theta^1) \) when \( \mu^1 = \mathbb{E}_{\theta^1}[\phi(X)] \).

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

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

\[ \ldots \text{can be written as:} \]

\[ \inf_{\mu \in \mathcal{M}} \{ A(\theta) + A^*(\mu) - \langle \theta, \mu \rangle \} = \inf_{\mu \in \mathcal{M}} D(\mu || \theta) = 0 \quad (12) \]

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

\[ \max_{\mu \in \mathcal{M}_F(G)} \{ \langle \mu, \theta \rangle - A^*_F(\mu) \} \quad (5) \]

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

- Mean field \( \equiv \) find the best approximation, in terms of this particular KL-divergence, to \( p_{\theta} \), over a family of “nice” distributions \( \mathcal{M}_F(G) \).

Naive Mean field for Ising Model

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

- Mean parameters for Ising: \( \mu_s = \mathbb{E}[X_s] = p(X_s = 1) \)

\[ \mu_{st} = \mathbb{E}[X_s X_t] = p(X_s = 1, X_t = 1), \text{ thus } \mu \in \mathbb{R}^{|V|+|E|}. \]

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

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

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

\[ -A^*(\mu) = -A^*_{F_0}(\mu) = \sum_{s \in V} H_s(\mu_s) \quad (13) \]

which are sum of unary entropy terms, very cheap.

- Moreover, polytope for \( \mathcal{M}_{F_0}(G) \) is also very simple, namely the hypercube \([0,1]^m\), with \( m = |V| \).
Naive Mean field for Ising Model

- We get variational lower bound problem

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

(14)

- 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.

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

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

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

(15)

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

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

- Consider simple two variable example \((X_1, X_2), X_i \in \{-1, +1\}\).
- Exponential family form
  \[
  p_\theta(x) \propto \exp(\theta_1 x_1 + \theta_2 x_2 + \theta_{12} x_1 x_2)
  \]
  having mean parameters \(\mu_i = \mathbb{E}[X_i]\) and \(\mu_{12} = \mathbb{E}[X_1 X_2]\).
- Impose constraint \(\mu_{12} = \mu_1 \mu_2\), we get mean field objective
  \[
  f(\mu_1, \mu_2; \theta) = \theta_{12} \mu_1 \mu_2 + \theta_1 \mu_1 + \theta_2 \mu_2 + H(\mu_1) + H(\mu_2)
  \]
  where \(H(\mu_i) = -\frac{1}{2}(1 + \mu_i) \log \frac{1}{2}(1 + \mu_i) - \frac{1}{2}(1 - \mu_i) \log \frac{1}{2}(1 - \mu_i)\)
  
  **Note that** \(p(X_i = +1) = \frac{1}{2}(1 + \mu_i)\)
- Consider sub-models of the form:
  \[
  (\theta_1, \theta_2, \theta_{12}) = \left(0, 0, \frac{1}{4} \log \frac{q}{1-q}\right) \triangleq \theta(q)
  \]
  where \(q \in (0, 1)\) is a parameter such that, for any \(q\) we have \(\mathbb{E}[X_i] = 0\).
  It turns out that in this form, we have \(q = p(X_1 = X_2)\).
- Is mean field objective in this case convex for all \(q\)?

Lack of Convexity example

- For \(q = 0.5\), objective \(f(\mu_1, \mu_2; \theta(0.5))\) has **global** maximum at \((\mu_1, \mu_2) = (0, 0)\) so mean field is exact and convex. This corresponds to \(p(X_1 = X_2) = 0\).
- When \(q\) gets small, \(f\) becomes non-convex, e.g., has multiple modes in figure.
key idea, set of sufficient statistics that yield efficient inference need
not be all independence. Could be a tree, or a chain, or a set of
trees/chains.

“structured” in general means that it is not a monolithic single variable,
but is a vector with some decomposability properties.

In Structured mean field, we exploit this and it again can be seen in our
variational framework.

Again, $\mathcal{I}(F)$ is set of suff. stats. corresponding to $F$, and we have
the corresponding mean vector $\mu(F) = (\mu_\alpha, \alpha \in \mathcal{I}(F))$.

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

Note also, $\mathcal{M}(F) \neq \mathcal{M}_F(G)$, their dimensions are entirely different.

Key thing: in mean field, $\mu(F) \in \mathcal{M}(F)$ and there is no real need to
mention the full $\mathcal{M}_F(G)$. Also, the dual $A^*_F$ depends on only $\mu(F)$ not
$\mu$ (the other values are derivations from entries within $\mu(F)$).

Other mean parameters $\mu_\beta$ for $\beta \in \mathcal{I} \setminus \mathcal{I}(F)$ do play a role in the value
of the mean field variational problem but their value is derivable from
values $\mu(F)$, thus we can express the $\mu_\beta$ in functional form based on
values $\mu(F)$.

Thus, for each $\beta \in \mathcal{I} \setminus \mathcal{I}(F)$, we set $\mu_\beta = g_\beta(\mu(F))$ for function $g_\beta$.

Example: mean field Ising, $\mu_{st} = g(\mu(F)) = \mu_s \mu_t$. 
Structured Mean Field

- The mean field optimization problem becomes

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

$$= \max_{\mu(F) \in \mathcal{M}(F)} \left\{ \sum_{\alpha \in \mathcal{I}(F)} \theta_\alpha \mu_\alpha + \sum_{\alpha \in \mathcal{I}^c(F)} \theta_\alpha g_\alpha(\mu(F)) - A^*_F(\mu(F)) \right\}$$

(20)

- With this, we can recover our sigmoid mean field coordinate update process by iterating fixed point equations of $f$, i.e.,

$$\frac{\partial f}{\partial \mu_\beta}(\mu(F)) = \theta_\beta + \sum_{\alpha \in \mathcal{I}(G) \setminus \mathcal{I}(F)} \theta_\alpha \frac{\partial g_\alpha}{\partial \mu_\beta}(\mu(F)) - \frac{\partial A^*_F}{\partial \mu_\beta}(\mu(F))$$

(21)

- Setting to zero, fix point condition is

$$\nabla A^*_F(\mu(F)) = \theta_\beta + \sum_{\alpha \in \mathcal{I}(G) \setminus \mathcal{I}(F)} \theta_\alpha \frac{\partial g_\alpha}{\partial \mu_\beta}(\mu(F))$$

(22)

- But recall, $\nabla A$ is the forward mapping, maps from mean to canonical, and $\nabla A^*$ does the reverse, thus letting $\gamma(F) = \nabla A(\mu(F))$, gives

$$\gamma_\beta(F) = \theta_\beta + \sum_{\alpha \in \mathcal{I}(G) \setminus \mathcal{I}(F)} \theta_\alpha \frac{\partial g_\alpha}{\partial \mu_\beta}(\mu(F))$$

(23)

- After this update, we might not have global consistency in the structured tractable portion so some fixup is needed. Would this be the case for simple (all independence) mean field?

- But since the structured tractable portion is (presumably) tractable, we can run message update to “fix” the tractable portion and regain global consistency.
Structured Mean Field

- Alternatively, we can transform back to mean parameters right away using cumulant $A^*$, of the form

$$\nabla A^*_F(\mu(F)) = \theta_\beta + \sum_{\alpha \in {I(G) \setminus I(F)}} \theta_\alpha \frac{\partial g_\alpha}{\partial \mu_\beta}(\mu(F)) \quad (24)$$

- But recall, $\nabla A$ is the forward mapping, maps from mean to canonical, and $\nabla A^*$ does the reverse, thus letting $\gamma(F) = \nabla A(\mu(F))$, gives

$$\gamma_\beta(F) = \theta_\beta + \sum_{\alpha \in {I(G) \setminus I(F)}} \theta_\alpha \frac{\partial g_\alpha}{\partial \mu_\beta}(\mu(F)) \quad (25)$$

- After this update, we might not have global consistency in the structured tractable portion so some fixup is needed. Would this be the case for simple (all independence) mean field?
- But since the structured tractable portion is (presumably) tractable, we can run message update to “fix” the tractable portion and regain global consistency.

We can also immediately transform back to mean parameters. I.e., we start with mean parameters, and transform them to new mean parameters in order to attempt to solve mean field variational problem.

- Tool to use is $A_f$ since $\nabla A_F(\gamma(F')) = \mu(F')$,
- thus we get update:

$$\mu_\beta(F) \leftarrow \frac{\partial A_F}{\partial \gamma_\beta} \left( \theta_\beta + \sum_{\alpha \in {I(G) \setminus I(F)}} \theta_\alpha \nabla g_\alpha(\mu(F)) \right) \quad (26)$$

- This generalizes our mean field coordinate ascent update from before, where $\frac{\partial A_F}{\partial \gamma_\beta}$ ended up being the sigmoid mapping.
Structured Mean Field Factorial HMMs

- This idea was developed using factorial HMMs.

![Diagram of interconnected circles representing HMM chains with dotted ellipses indicating induced dependencies]

- While each HMM chain is simple (it is only a chain), the common observation induces a dependence between each. Thus, if there are $M$ chains, we have a clique of size $M$.

The induced dependencies (cliques as dotted ellipses)

![Diagram with dotted ellipses around clusters of HMM chains]

- Tree width of this model is $M$
- Thus, if $r$ states per chain, then complexity $r^{M+1}$.
Structured Mean Field Factorial HMMs

- A “natural” choice of approximating distribution is a set of coupled chains. Why is this natural? Perhaps for computational reasons only.

- With this independent chains case, we have that 
  \[ g_\beta(\mu(F)) = \prod_i f_i(\mu_i(F)) \]
  decomposable, so this is easier to work with

- forward backward procedure on a chain is necessary to recover chain consistency.

Variational Problem

Original variational representation of log partition function

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

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

- Set \( M \leftarrow \mathbb{I}_E(G) \) (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 \( M \leftarrow \mathcal{L}(\phi, \Phi) \) and set
  \( -A^*(\mu) \leftarrow H_{\text{ep}}(\tau, \tilde{\tau}) \) to get expectation propagation.

- Mean field (from variational approximation perspective) is
  \[ A(\theta) \geq \max_{\mu \in \mathcal{M}_F(G)} \{ \langle \mu, \theta \rangle - A^*_F(\mu) \} \]
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.
Convex Relaxations and Upper Bounds - Relaxed Entropy

- Given $\mu \in \mathcal{M}$, $\mu(F) \in \mathcal{M}(F)$ projects from $\mathcal{I}$ to $\mathcal{I}(F)$.
- 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 3

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

$$A^*(\mu(F)) \leq A^*(\mu) \quad (29)$$

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

\[ A^*(\mu) = \sup_{\theta \in \mathbb{R}^d} \{ \langle \mu, \theta \rangle - A(\theta) \} \] (30)

- Dual problem in sub-graph case.

\[ A^*(\mu(F)) = \sup_{\theta(F) \in \mathbb{R}^{d(F)}} \{ \langle \mu(F), \theta(F) \rangle - A(\theta(F)) \} \] (31)

- Dual problem in sub-graph case — alternate expression

\[ A^*(\mu(F)) = \sup_{\theta \in \mathbb{R}^d} \{ \langle \mu, \theta \rangle - A(\theta) \} \] (32)

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

- Thus, \( A^*(\mu) \geq A^*(\mu(F)) \).

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

- Distribution over tractable structures \( \rho \in \mathbb{R}^{[\mathcal{D}]} \), 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

\[ H(\mu) \leq \mathbb{E}_\rho[H(\mu(F))] = \sum_{F \in \mathcal{D}} \rho(F)H(\mu(F)) \] (33)

- 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:

  \[
  \mathcal{L}(G; \mathcal{D}) = \left\{ \tau \in \mathbb{R}^d | \tau(F) \in \mathcal{M}(F) \ \forall F \in \mathcal{D} \right\} \quad (34)
  \]

  \[
  = \bigcap_{F \in \mathcal{D}} \mathcal{M}(F) \quad (35)
  \]

- 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 $\mathcal{M}$, we get a new variational approximation to the cumulant function.

  \[
  B_{\mathcal{D}}(\theta; \rho) \triangleq \sup_{\tau \in \mathcal{L}(G; \mathcal{D})} \left\{ \langle \tau, \theta \rangle + \sum_{F \in \mathcal{D}} \rho(F)H(\tau(F)) \right\} \quad (36)
  \]

- 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 $\mathcal{M}(G)$

- Thus $B_{\mathcal{D}}(\theta; \rho)$ is convex, has a global optimal solution, it approximates $A(\theta)$, and best of all is an upper bound, $A(\theta) \leq B_{\mathcal{D}}(\theta; \rho)$
We can get convex upper bounds in the tree case, and a new style of sum-product algorithm.

Consider MRF again

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

(37)

Let \( \mathcal{T} \) be a set of all spanning trees \( T \) of \( G \), and let \( \rho \) be a distribution over them, \( \sum_T \rho(T) = 1 \).

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.,

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

(38)

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 \( \mathcal{T} \).

The expression becomes

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

(39)

Note sum is over all \( E \) and terms are weighted by probability.
Tree-reweighted sum-product and Bethe

- We also need outer bound.
- 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 $\mathbb{L}(T)$, the pseudo-marginals (which 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 $T$, this asks 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 $\mathbb{L}(G)$.
- 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) bound on entropy.

Theorem 4 (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):

$$B_\Sigma(\theta; \rho) = \max_{\tau \in \mathbb{L}(G)} \left\{ \langle \tau, \theta \rangle + \sum_{s \in V} H_s(\tau_s) - \sum_{(s,t) \in E} \rho_{st} I_{st}(\tau_{st}) \right\} \quad (40)$$

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

...
Theorem 4 (Tree-Reweighted Bethe and Sum-Product)

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

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

(41)

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).

Note that if $\rho_{st} = 1$, then we get back standard LBP and Bethe approximation.

Thus, this is a true convex generalization.

However, if $\rho_{st} = 1$, for all spanning trees, it must mean that $G = T$ and we get back standard tree-based message passing we saw in lecture 2!!
**Sources for Today’s Lecture**

- this material comes from the Wainwright and Jordan book.