EE512A – Advanced Inference in Graphical Models

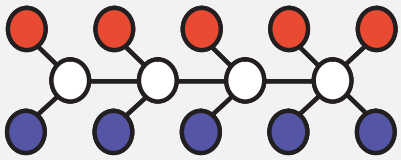
— Fall Quarter, Lecture 18 —

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

Dec 1st, 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 through 5 in our book. Read chapter 7.
- Also read chapter 8 (integer programming, although we probably won't cover that chapter in class unfortunately).
- Also should have 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 (Dec 3rd) night, 11:45pm. Final project proposal final progress report (one page max).
- Update: For status update, final writeup, and talk, use notation as close as possible to that used in class!

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): Structured mean field, Convex relaxations and upper bounds, tree reweighted case
- L19 (12/3):
- Final Presentations: (12/10):

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

Conjugate Duality, Maximum Likelihood, Negative Entropy

Theorem 18.2.3 (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:*

$$A^*(\mu) = \sup_{\theta \in \Omega} \left(\langle \theta, \mu \rangle - A(\theta) \right) = \begin{cases} -H(p_{\theta(\mu)}) & \text{if } \mu \in \mathcal{M}^\circ \\ +\infty & \text{if } \mu \notin \overline{\mathcal{M}} \end{cases} \tag{18.3}$$

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

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \left\{ \langle \theta, \mu \rangle - A^*(\mu) \right\} \tag{18.4}$$

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

$$\mu = \int_{\mathsf{D}_X} \phi(x) p_\theta(x) \nu(dx) = \mathbb{E}_\theta[\phi(X)] = \nabla A(\theta) \tag{18.5}$$

Variational Approach Amenable to Approximation
Variational Approximations we cover

- Original variational representation of log partition function

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{18.1}$$

where dual takes form:

$$A^*(\mu) = \sup_{\theta \in \Omega} (\langle \theta, \mu \rangle - A(\theta)) = \begin{cases} -H(p_{\theta(\mu)}) & \text{if } \mu \in \mathcal{M}^\circ \\ +\infty & \text{if } \mu \notin \overline{\mathcal{M}} \end{cases} \tag{18.2}$$

- 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 $H_{\text{app}} = \sum_{g \in E} c(g) H_g(\tau_g)$ (via Mobius) to get Kikuchi variational approximation, message passing on hypergraphs.
3. Partition τ 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.

Logistics Review

Variational Approach Amenable to Approximation
Variational Approximations we cover

- Original variational representation of log partition function

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{18.1}$$

where dual takes form:

$$A^*(\mu) = \sup_{\theta \in \Omega} (\langle \theta, \mu \rangle - A(\theta)) = \begin{cases} -H(p_{\theta(\mu)}) & \text{if } \mu \in \mathcal{M}^\circ \\ +\infty & \text{if } \mu \notin \overline{\mathcal{M}} \end{cases} \tag{18.2}$$

- 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 $H_{\text{app}} = \sum_{g \in E} c(g) H_g(\tau_g)$ (via Mobius) to get Kikuchi variational approximation, message passing on hypergraphs.

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, clusters, or many other structures as well).
- Can also be done for Gaussian mixture and other distributions.
- 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 $\mathcal{M}$.
- 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 $A(\theta)$, all else (i.e., the entropy) being equal.
- 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 of the optimization problem is often lost still, however, in the general case (due to the inner bound).
- Thus, in mean field, we will get a lower bound on $A(\theta)$ but not a convex procedure to find it (both good and bad news).

Tractable Families (for mean field approach)

- 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, canonical F-respecting parameters are of the form:

$$\mathbb{R}^{|\mathcal{I}|} \ni \Omega(F) \triangleq \{\theta \in \Omega | \theta_\alpha = 0 \ \ \forall \alpha \in \mathcal{I} \setminus \mathcal{I}(F)\} \subseteq \Omega. \tag{18.14}$$

  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.

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 F-respecting models. I.e.,

$$\mathcal{M}_F(G; \phi) = \left\{ \mu \in \mathbb{R}^d | \mu = \mathbb{E}_\theta[\phi(x)] \text{ for some } \theta \in \Omega(F) \right\} \tag{18.18}$$

- Therefore, since $\theta \in \Omega(F) \subseteq \Omega$, we have that

$$\mathcal{M}_F^\circ(G; \phi) \subseteq \mathcal{M}^\circ(G; \phi) \tag{18.19}$$

  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 $M^\circ(G) = M^\circ(G; \phi)$

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

$$A(\theta) \geq \max_{\mu \in \mathcal{M}_F(G)} \{ \langle \mu, \theta \rangle - A_F^*(\mu) \} \triangleq A_{\text{MF}}(\theta) \tag{18.23}$$

  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 $\mu \in \mathcal{M}_F(G)$ and since $\mathcal{M}_F(G) \subseteq \mathcal{M}(G)$, $A_F^*(\mu)$ is not an approximation, rather it is just easy to compute.

Mean field, KL-Divergence, Exponential Model Families

- Thus, solving the mean-field variational problem (see Eqn. (18.23)) of:

$$\max_{\mu \in \mathcal{M}_F(G)} \{ \langle \mu, \theta \rangle - A_F^*(\mu) \} = \max_{\mu \in \mathcal{M}_F(G)} \{ \langle \mu, \theta \rangle - A^*(\mu) \} \tag{18.34}$$

  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: optimization

- We get variational lower bound problem

$$A(\theta) \geq \max_{(\mu_1,\ldots,\mu_m)\in[0,1]^m} \left\{ \sum_{s\in V} \theta_s \mu_s + \sum_{(s,t)\in E} \theta_{st}\mu_s\mu_t + \sum_{s\in V} H_s(\mu_s) \right\} \tag{18.35}$$

- 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.
- We have a non-convex problem, so while it is a bound, it might be hard to get as tight as possible.
- 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.
- Fortunately, each coordinate optimization is concave!

Structured Mean Field

- 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.
- We first see a nice way that we can use fixed points of the mean field primal/dual equations to derive a general form of the mean field update.

Structured Mean Field

- Again, $\mathcal{I}(F)$ is set of suff. stats. corresponding to F, and we have corresponding mean vector $\mu(F) = (\mu_\alpha, \alpha \in \mathcal{I}(F))$.
- Define new quantity $\mathcal{M}(F)$, the 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 $M_F(G)$. Also, the dual A^*_F depends on only $\mu(F)$ not μ (the other values are derivations from entries within $\mu(F)$).
- Other mean parameters μ_β 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 μ_β 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_β.
- Ex: mean field Ising, edges $(s,t) \in E$, get $\mu_{st} = g_{st}(\mu(F)) = \mu_s \mu_t$.

Structured Mean Field

- The mean field optimization problem becomes

$$\max_{\mu \in \mathcal{M}_F(G)} \{\langle \mu, \theta \rangle - A^*_F(\mu)\} \tag{18.1}$$

$$= \max_{\mu(F) \in \mathcal{M}(F)} \underbrace{\left\{ \sum_{\alpha \in \mathcal{I}(F)} \theta_\alpha \mu_\alpha + \sum_{\alpha \in \mathcal{I}^c(F)} \theta_\alpha g_\alpha(\mu(F)) - A^*_F(\mu(F)) \right\}}_{f(\mu(F))} \tag{18.2}$$

- With this, we can recover our sigmoid mean field coordinate update process by iterating fixed point equations of f, i.e., for $\beta \in \mathcal{I}(F)$,

$$\frac{\partial f}{\partial \mu_\beta}(\mu(F)) = \theta_\beta + \sum_{\alpha \in \mathcal{I}(G) \setminus \mathcal{I}(F)} \theta_\alpha \frac{\partial g_\alpha}{\partial \mu_\beta}(\mu(F)) - \frac{\partial A^*_F}{\partial \mu_\beta}(\mu(F)) \tag{18.3}$$

Structured Mean Field

- Setting this to zero, and then aggregating/concatenating over $\beta \in \mathcal{I}(F)$, vector fix point condition is:

$$\nabla A_F^*(\mu(F)) = \theta + \sum_{\alpha \in \mathcal{I}(G) \setminus \mathcal{I}(F)} \theta_\alpha \nabla g_\alpha(\mu(F)) \tag{18.4}$$

- ∇A is the forward mapping, maps from canonical to mean parameters, and ∇A^* does the reverse. Hence, naming $\gamma(F) = \nabla A(\mu(F))$, gives a parameter update equation for $\beta \in \mathcal{I}(F)$

$$\gamma_\beta(F) \leftarrow \theta_\beta + \sum_{\alpha \in \mathcal{I}(G) \setminus \mathcal{I}(F)} \theta_\alpha \frac{\partial g_\alpha}{\partial \mu_\beta}(\mu(F)) \tag{18.5}$$

- Above is the mean field update, mapping from canonical parameters (θ_β, and θ_α for $\alpha \in \mathcal{I}(G) \setminus \mathcal{I}(F)$) and using the mean parameters $\mu(F)$ to new updated canonical parameters $\gamma_\beta(F)$ for $\beta \in \mathcal{I}(F)$). It is to be repeated over and over.

Structured Mean Field

- After each update of Eqn. (18.5), a mean parameter, say $\mu(F)_\delta$, that depends on any of the updated canonical parameter also needs to be updated before doing the next update.
- Since we're using a tractable sub-structure F, we can then update the out-of-date mean parameters using any exact inference algorithm (e.g., junction tree, possible since sub-structure is tractable), and then repeat Eqn. (18.5).

Structured Mean Field

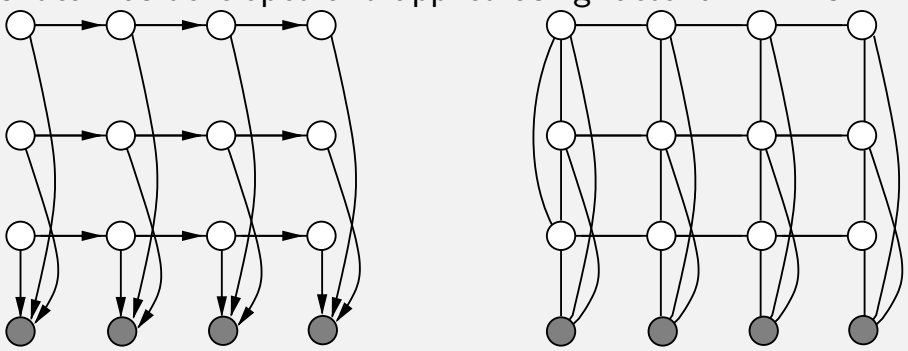
- Alternatively, we can transform back to mean parameters right away using ∇A is the forward mapping, maping from canonical to mean.
- I.e., we can derive a mean field mean parameter to mean parameter update equation using A_F since $\nabla A_F(\gamma(F)) = \mu(F)$,
- We get update, for $\beta \in \mathcal{I}(F)$:

$$\mu_\beta(F) \leftarrow \frac{\partial A_F}{\partial \gamma_\beta}\left(\theta_\beta + \sum_{\alpha \in \mathcal{I}(G)\setminus\mathcal{I}(F)} \theta_\alpha \nabla g_\alpha(\mu(F))\right) \tag{18.6}$$

- This generalizes our mean field coordinate ascent update from before, where in that case we would get $\frac{\partial A_F}{\partial \gamma_\beta}$ as being the sigmoid mapping.
- But here, we can use this for any tractable substructure (e.g., trees or chains or collections thereof).

Structured Mean Field Factorial HMMs

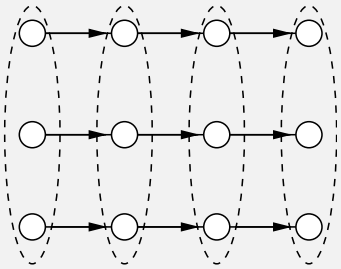
- This idea was developed and applied using factorial HMMs.



- Graph consists of M 1st-order Markov chains $x^i_{1:T}$ for $i \in [M]$, coupled together at each time via factor $p(\bar{y}_t | x^1_t, x^2_t, \ldots, x^M_t)$.
- While each HMM chain is simple (it is only a chain, so a 1-tree), the common observation induces a dependence between each. Thus, given M chains, have a clique of size M (e.g., after moralization, on right)
- After moralization, covering hypergraph consists of tractable sub-substructure hyperedges $F = \left\{\left\{x^i_t, x^i_{t+1}\right\} : i \in [M], t \in [T]\right\}$ and remaining structure $E \setminus F = \left\{\left\{x^1_t, x^2_t, \ldots, x^M_t\right\} : t \in [T]\right\}$.

Structured Mean Field Factorial HMMs

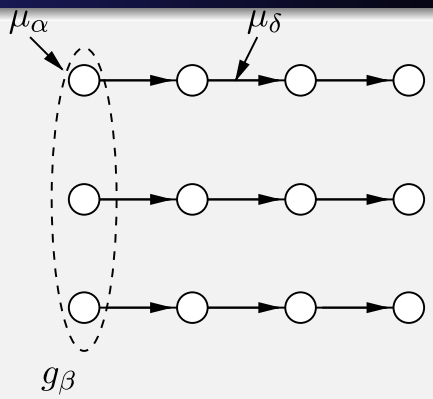
- The induced dependencies (cliques as dotted ellipses)



- Tree width of this model is? M
- Thus, if r states per chain, then exact inference complexity r^{M+1}.
- Each $\beta \in \mathcal{I}(F)$ corresponds to one of the Markov chain edges in one of the M Markov chains, each soting $O(r^2)$.
- Each $\beta \in \mathcal{I} \setminus \mathcal{I}(F)$ corresponds to one of the size M cliques (dotted ellipses above) corresponding to the v-structure moralizations, each costing $O(r^M)$.

Structured Mean Field Factorial HMMs

- A "natural" choice of approximating distribution is a set of coupled chains, natural, perhaps primarily for computational reasons.



- Under this independent chains case, we have that for each $\beta \in \mathcal{I} \setminus \mathcal{I}(F)$, derivable functions have form $g_\beta(\mu(F)) = \prod_{i=1}^{M} f_i(\{\mu_i(F)\})$, for some functions f_i. This is fully factored, so is easy to work with, maintains separate chains.
- Each update of form Eqn. (18.5) updates parameters for $\beta \in \mathcal{I}(F)$, corresponds to all edges of all M Markov chains.
- To recover mean parameters (or do Eqn. (18.6)), need only forward-backward procedure on each chain separately, $O(MTr^2)$.

Convex Relaxations and Upper Bounds

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{18.7}$$

- Other than mean field (which gives lower bound on $A(\theta)$), none of the other approximation methods have been anything other than approximation methods.
- What about upper bounds?
- 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

- Recall sufficient stats $\phi = (\phi_\alpha, \alpha \in \mathcal{I})$ and canonical parameters $\theta = (\theta_\alpha, \alpha \in \mathcal{I})$.
- In general, inference (computing mean parameters) starting from canonical 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 $\mathfrak{D}$ be a set of subfamilies that are tractable.
- I.e., $\mathfrak{D}$ might be all spanning trees of G, or some subset of spanning trees that we like.
- As before, $\mathcal{I}(F) \subseteq \mathcal{I}$ are the subset of indices of the suff. stats. that abide by F, and $|\mathcal{I}(F)| = d(F) < d = |\mathcal{I}|$ suff. stats.
- As before, $\mathcal{M}(F)$ is set of realizable mean parameters associated with F, and $\mu(F) \in \mathcal{M}(F)$. Thus, $\mathcal{M}(F) \subseteq \mathbb{R}^{|\mathcal{I}(F)|}$, and

$$\mathcal{M}(F) = \left\{ \mu \in \mathbb{R}^{|\mathcal{I}(F)|} | \exists p \text{ s.t. } \mu_\alpha = \mathbb{E}_p[\phi_\alpha(X)] \; \forall \alpha \in \mathcal{I}(F) \right\} \tag{18.8}$$

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

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.

Convex Relaxations and Upper Bounds - Relaxed Entropy

Proposition 18.4.1 (Maximum Entropy Bounds)

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

$$A^*(\mu(F)) \leq A^*(\mu) \tag{18.9}$$

or alternatively stated, $H(\mu(F)) \geq H(\mu)$, entropy of projection is higher.

- Intuition: $H(\mu) = H(p_\mu)$ is the entropy of the exponential family model with mean parameters μ.
- 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 $\mu = \mathbb{E}_{p_\theta}[\phi(X)]$.
- Fewer constraints when forming $\mu(F)$ (see Eqn. (18.8)), so entropy in corresponding maxent problem can only, if anything, get larger.
- Thus, $H(\mu(F)) \geq H(\mu)$.

Convex Relaxations and Upper Bounds - Relaxed Entropy

Proof.

- Dual problem

$$A^*(\mu) = \sup_{\theta \in \mathbb{R}^d} \{\langle \mu, \theta \rangle - A(\theta)\} \tag{18.10}$$

- Dual problem in sub-graph case.

$$A^*(\mu(F)) = \sup_{\theta(F) \in \mathbb{R}^{d(F)}} \{\langle \mu(F), \theta(F) \rangle - A(\theta(F))\} \tag{18.11}$$

- Dual problem in sub-graph case — alternate expression

$$A^*(\mu(F)) = \sup_{\substack{\theta \in \mathbb{R}^d \\ \theta_\alpha = 0 \, \forall \alpha \notin \mathcal{I}(F)}} \{\langle \mu, \theta \rangle - A(\theta)\} \tag{18.12}$$

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

Convex Relaxations and Upper Bounds - Relaxed Entropy

- Note that the upper bound is true for each $F \in \mathfrak{D}$, and thus would be true for mixtures of different $F \in \mathfrak{D}$.
- We can form a distribution over tractable structures, i.e., $\rho \in \mathbb{R}^{|\mathfrak{D}|}$, i.e., $\rho(F) \geq 0$ for $F \in \mathfrak{D}$ and $\sum_{F \in \mathfrak{D}} \rho(F) = 1$
- Convex combination over $F \in \mathfrak{D}$, gives more general upper bound

$$H(\mu) \leq \mathbb{E}_\rho[H(\mu(F))] = \sum_{F \in \mathfrak{D}} \rho(F) H(\mu(F)) \tag{18.13}$$

- This will be our convexified upper bound on entropy (lower bound on the dual).
- compared to mean field, we are not choosing only one structure, but many of them, and mixing them together in a certain way.
- This so far gives us an upper bound on $A(\theta)$, but we still need an outer bound. The combination will give us our uppper bound on $A(\theta)$.

Convex Relaxations and Upper Bounds - Outer bound

- When we form mixture of entropies (which really are duals), we make sure any given $\mu(F)$ can be evaluated for any dual (i.e., each component can properly evaluate any possible $\mu(F)$).
- Logical constraint: make sure any $\mu(F)$ works for all components.
- Constraint set as follows:

$$\mathcal{L}(G;\mathfrak{D}) = \left\{\tau \in \mathbb{R}^d | \tau(F) \in \mathcal{M}(F) \;\; \forall F \in \mathfrak{D}\right\} \tag{18.14}$$

- Note this is an outer bound i.e., $\mathcal{L}(G;\mathfrak{D}) \supseteq \mathcal{M}(G)$ since any member of $\mathcal{M}(G)$ (any valid mean parameter for G) must also be a member of any $\mathcal{M}(F)$.
- Also note, $\mathcal{L}(G;\mathfrak{D})$ 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.

$$B_{\mathfrak{D}}(\theta;\rho) \stackrel{\Delta}{=} \sup_{\tau \in \mathcal{L}(G;\mathfrak{D})} \left\{ \langle \tau, \theta \rangle + \sum_{F \in \mathfrak{D}} \rho(F) H(\tau(F)) \right\} \tag{18.15}$$

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

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

$$p_\theta(x) \propto \exp\left\{\sum_{s\in V}\theta_s(x_s) + \sum_{(s,t)\in E}\theta_{st}(x_s,x_t)\right\} \tag{18.16}$$

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

$$H(\mu(T)) = \sum_{s\in V}H_s(\mu_s) - \sum_{(s,t)\in E(T)} I_{st}(\mu_{st}) \tag{18.17}$$

- We want to use this to see what happens when we take the expected value w.r.t. distribution ρ.

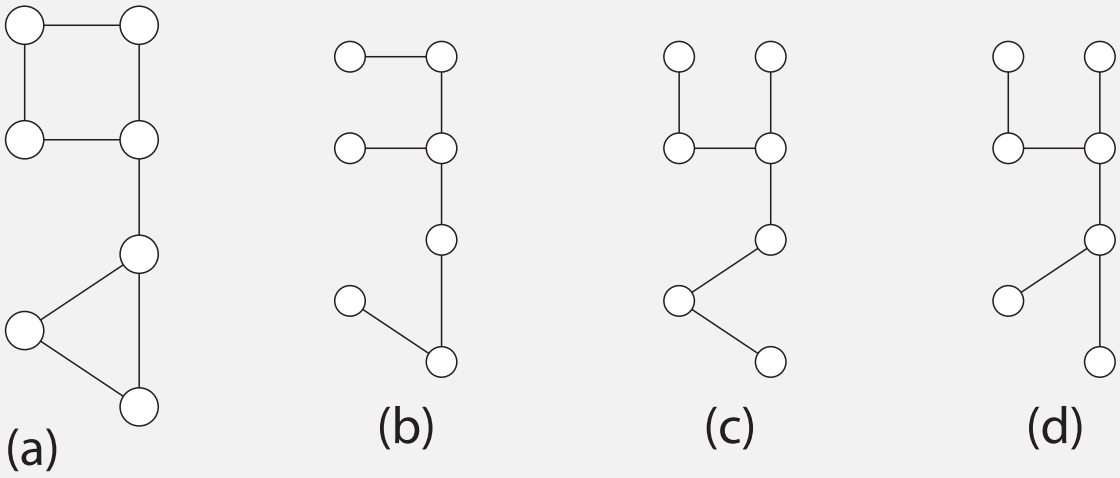
Tree-reweighted sum-product and Bethe

- Every tree is spanning, all tress have all nodes, so the probability, according to ρ of 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 $\mathfrak{T}$.
- The expression becomes

$$H(\mu) \leq \sum_{s\in V}H_s(\mu_s) - \sum_{(s,t)\in E}\rho_{st}I_{st}(\mu_{st}) \tag{18.18}$$

  Note right hand sum is over all E (not just a given spanning tree) and terms are weighted by probability of the given edge ρ_{st}.
- ρ_{st} is edge appearance probability, $\rho = (\rho_{st}, (s,t)\in E)$ is spanning tree polytope.

Edge appearance probabilities example



- (a) a graph $G = (V, E)$ with $m = |V| = 7$
- (b), (c), and (d) various spanning trees, each with probability 1/3.
- What are the edge appearance probabilities ρ_{st}?

Tree-reweighted sum-product and Bethe

- We also need outer bound on $\mathcal{M}$.
- For discrete case $\mathcal{M} = \mathbb{M}(G)$ is marginal polytope.
- $\mathbb{M}(T)$ is marginal polytope for tree, and for a tree is the same as $\mathbb{L}(T)$, the locally consistent pseudo-marginals (which recall are marginals for a tree).
- Thus, $\mu(T) \in \mathbb{M}(T)$ 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 $\mu(T) \in \mathbb{M}(T)$ for all $T \in \mathfrak{T}$, this is identical to asking for local marginalization on every edge of G.
- Thus, in this case $\mathcal{L}(G; \mathfrak{I})$ 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 based) bound on entropy.

Tree-reweighted sum-product and Bethe

Theorem 18.5.1 (Tree-Reweighted Bethe and Sum-Product)

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

$$B_{\mathfrak{T}}(\theta;\rho) = \max_{\tau \in \mathbb{L}(G)} \left\{ \langle \tau, \theta \rangle + \sum_{s \in V} H_s(\tau_s) - \sum_{(s,t) \in E} \rho_{st} I_{st}(\tau_{st}) \right\} \tag{18.19}$$

$$\geq A(\theta) \tag{18.20}$$

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

. . .

Tree-reweighted sum-product and Bethe

Theorem 18.5.1 (Tree-Reweighted Bethe and Sum-Product)

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

$$M_{t \to s}(x_s) \leftarrow \kappa \sum_{x'_t \in \mathcal{X}_t} \varphi_{st}(x_s, x'_t) \frac{\prod_{v \in N(t) \setminus \{s\}} [M_{v \to t}(x'_t)]^{\rho_{vt}}}{[M_{s \to t}(x'_t)]^{(1-\rho_{ts})}} \tag{18.21}$$

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 assumptions given in (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.
- 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!!
- Thus, this is a true convex generalization, when $\rho_{st} < 1$ for many s,t.
- Note that $\rho = (\rho_{st}, (s,t) \in E)$ must live in the "spanning tree polytope" $\subseteq \mathbb{R}_+^E$, i.e., a convex combination of vertices consisting of characteristic (indicator) functions of spanning trees (see example earlier). I.e., Let $\mathfrak{T}$ be the set of all spanning trees, and $\mathbf{1}_T \in \{0,1\}^E$ be the characteristic vector of $T \in \mathfrak{T}$. Then we must have that

$$\rho \in \mathrm{conv}(\{\mathbf{1}_T : T \in \mathfrak{T}\}) \tag{18.22}$$

  where $\mathrm{conv}(\cdot)$ is the convex hull of its argument.

More on spanning tree polytope

- Spanning tree polytope takes the form

$$\rho \in \mathrm{conv}(\{\mathbf{1}_T : T \in \mathfrak{T}\}) \tag{18.23}$$

  where $\mathfrak{T}$ is set of all spanning trees.
- Consider graphic matroid on $G = (V,E)$ with rank function $r(A)$ for any $A \subseteq E$.
- Then A is a spanning tree iff $r(A) = |A|$ and $|A| = m - 1$.
- Consider polytopes:

$$P_r = \left\{x \in \mathbb{R}_+^E : x(A) \leq r(A), \forall A \subseteq E\right\} \tag{18.24}$$

$$B_r = P_r \cap \left\{x \in \mathbb{R}_+^E : x(E) = r(E)\right\} \tag{18.25}$$

- Then if T is a spanning tree, $\mathbf{1}_T \in B_r$, and $B_r = \mathrm{conv}(\{\mathbf{1}_T : T \in \mathfrak{T}\})$.
- Edmonds showed that a simple fast greedy procedure will maximize a linear function over this polytope, and this can be useful for finding good points in the spanning tree polytope.

Tree-reweighted sum-product: convex vs. upper bound

- 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 verse.

Tree-reweighted sum-product fixed point

The fixed point we ultimately reach has following form:

$$\tau_s^*(x_s) = \kappa \exp\{\theta_s(x_s)\} \prod_{v \in N(s)} [M^*_{v \to s}(x_s)]^{\rho_{vs}} \tag{18.26}$$

$$\tau_{st}^*(x_s, x_t) = \kappa \varphi_{st}(x_s, x_t) \frac{\prod_{v \in N(s)\setminus t} [M^*_{vs}(x_s)]^{\rho_{vs}} \prod_{v \in N(t)\setminus s} [M^*_{vt}(x_t)]^{\rho_{vt}}}{[M^*_{ts}(x_s)]^{(1-\rho_{st})} [M^*_{st}(x_t)]^{(1-\rho_{ts})}} \tag{18.27}$$

with $\varphi_{st}(x_s, x_t) = \exp\left\{\frac{1}{\rho_{st}}\theta_{st}(x_s, x_t) + \theta_s(x_s) + \theta_t(x_t)\right\}$ where the $*$ versions are the final (convergent) messages.

- In practice: damping of messages M appears in practice to help reach convergence, where each new message is a convex mixture of the previous version of itself and the new message according to the equations.

hypertree-reweighted sum-product

- 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. $\mathfrak{T}(t)$.
- Then we get the same form of bounds

$$H(\mu) \leq E_\rho[H(\mu(T))] = \sum_{T \in \mathfrak{T}(t)} \rho(T) H(\mu(T)) \tag{18.28}$$

  but here T is over all valid k-trees.
- This leads to a convexified Kikuchi variational problem

$$A(\theta) \leq B_{\mathfrak{B}(t)}(\theta; \rho) = \max_{\tau \in \mathbb{L}(G)} \{\langle \tau, \theta \rangle + \mathbb{E}_\rho[H(\tau(T))]\} \tag{18.29}$$

  same form (but different than) before.
- Optimizing ρ over hypertree polytope is hard, unfortunately.

Reweighted EP

- Other variational variants have convexified version.
- Convexified forms of EP

$$H_{\text{ep}}(\tau, \tilde{\tau}; \rho) = H(\tau) + \sum_{\ell=1}^{d_I} \rho(\ell)[H(\tau, \tilde{\tau}^\ell) - H(\tau)] \tag{18.30}$$

  where $\sum_\ell \rho(\ell) = 1$.
- In this case, reweighted entropy is concave!
- Lagrangian formulation leads to solutions that are a form of "reweighted" EP, ideas which also are sometimes called "power EP" (blending the above reweighted sum-product ideas and EP).

Other variants

- Why only trees? There could be other tractable families (e.g., perhaps planar graphs, or 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.
- Other examples from book:
  - Use of Gaussian continuous entropy as an upper bound and a covariance-based outer bound of $\mathcal{M}$.
  - use of conditional entropy, various forms of use of polyhedral approximations.
- This is still an active research area!

Variational Approach Amenable to Approximation
Variational Approximations we cover

- Original variational representation of log partition function

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{18.1}$$

where dual takes form:

$$A^*(\mu) = \sup_{\theta \in \Omega} (\langle \theta, \mu \rangle - A(\theta)) = \begin{cases} -H(p_{\theta(\mu)}) & \text{if } \mu \in \mathcal{M}^\circ \\ +\infty & \text{if } \mu \notin \overline{\mathcal{M}} \end{cases} \tag{18.2}$$

- 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 $H_{\text{app}} = \sum_{g \in E} c(g) H_g(\tau_g)$ (via Mobius) to get Kikuchi variational approximation, message passing on hypergraphs.

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`