EE512A – Advanced Inference in Graphical Models

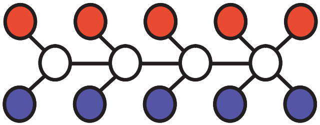
— Fall Quarter, Lecture 18 —

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:*

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

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

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

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

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

Variational Approach Amenable to Approximation

- Original variational representation of log partition function

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

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.

Variational Approximations we cover

1. Set $\mathcal{M} \leftarrow \mathbb{L}$ and $-A^*(\mu) \leftarrow H_{\text{Bethe}}(\tau)$ to get Bethe variational approximation, LBP fixed point.
2. Set $\mathcal{M} \leftarrow \mathbb{L}_t(G)$ (hypergraph marginal polytope), $-A^*(\mu) \leftarrow H_{\text{app}}(\tau)$ where $H_{\text{app}} = \sum_{g \in E} c(g) H_g(\tau_g)$ (via Möbius) to get Kikuchi variational approximation, message passing on hypergraphs.
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.
4. Mean field (from variational perspective) is (with $\mathcal{M}_F(G) \subseteq \mathcal{M}$) **l.b.**:

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

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 $A^*(\mu) = \sup_{\theta\in\Omega}(\langle\theta,\mu\rangle - A(\theta))$ is intractable or unavailable, but key idea is that if $\mu \in \mathcal{M}_F(G)$ it will be possible to compute easily.
- Thus, goal of mean field (from variational approximation perspective) is to form $A_{\mathsf{MF}}(\theta)$ where:

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

  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. (**??**)) 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 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.

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.

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 $\mu(F) = (\mu_\alpha, \alpha \in \mathcal{I}(F))$.

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, $\mathcal{M}(F) \subseteq \mathbb{R}^{|\mathcal{I}(F)|}$.

$M_F(\sigma)$

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.

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 $\mu(F)$.

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 $\mu(F)$.

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.

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

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

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

$\mathcal{I}^c(F) = \mathcal{I} \setminus \mathcal{I}(F)$

Structured Mean Field

- The mean field optimization problem becomes

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

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

- 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:

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

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 $\beta \in \mathcal{I}(F)$

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

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.

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.

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 $\nabla A_F(\gamma(F)) = \mu(F)$,

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 $\beta \in \mathcal{I}(F)$:

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

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.

Structured Mean Field

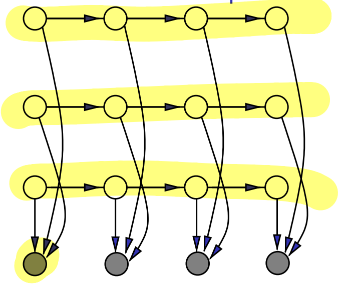
- Alternatively, we can transform back to mean parameters right away using ∇A is the forward mapping, maping from canonical to mean.
- I.e., we can derive a mean field mean parameter to mean parameter update equation using A_F since $\nabla A_F(\gamma(F)) = \mu(F)$,
- We get update, for $\beta \in \mathcal{I}(F)$:

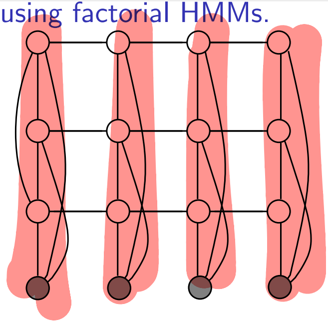
$$\mu_\beta(F) \leftarrow \frac{\partial A_F}{\partial \gamma_\beta}\left(\theta_\beta + \sum_{\alpha \in \mathcal{I}(G) \setminus \mathcal{I}(F)} \theta_\alpha \nabla g_\alpha(\mu(F))\right) \tag{18.6}$$

- 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

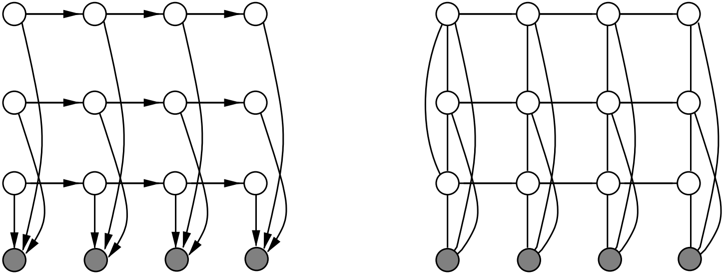
- This idea was developed and applied using factorial HMMs.





Structured Mean Field Factorial HMMs

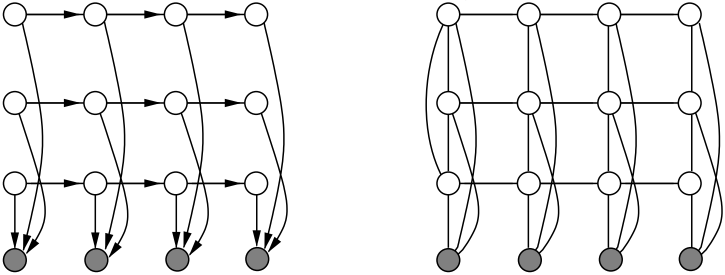
- This idea was developed and applied using factorial HMMs.



- Graph consists of M 1st-order Markov chains $x^i_{1:T}$ for $i \in [M]$, coupled together at each time via factor $p(\bar{y}_t | x^1_t, x^2_t, \ldots, x^M_t)$.

Structured Mean Field Factorial HMMs

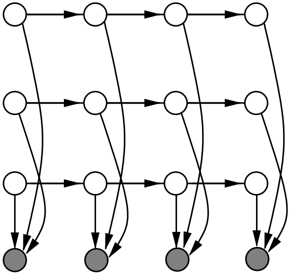
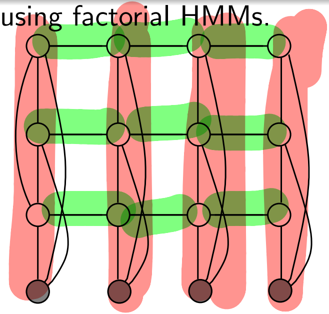
- This idea was developed and applied using factorial HMMs.



- 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)

Structured Mean Field Factorial HMMs

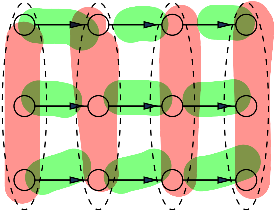
- This idea was developed and applied using factorial HMMs.

- Graph consists of M 1st-order Markov chains $x^i_{1:T}$ for $i \in [M]$, coupled together at each time via factor $p(\bar{y}_t | x^1_t, x^2_t, \dots, x^M_t)$.
- 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 $F = \{\{x^i_t, x^i_{t+1}\} : i \in [M], t \in [T]\}$ and remaining structure $E \setminus F = \{\{x^1_t, x^2_t, \dots, x^M_t\} : t \in [T]\}$.

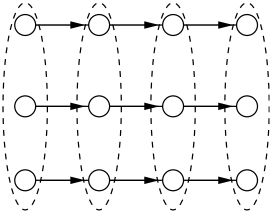
Structured Mean Field Factorial HMMs

- The induced dependencies (cliques as dotted ellipses)



Structured Mean Field Factorial HMMs

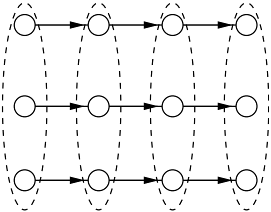
- The induced dependencies (cliques as dotted ellipses)



- Tree width of this model is?

Structured Mean Field Factorial HMMs

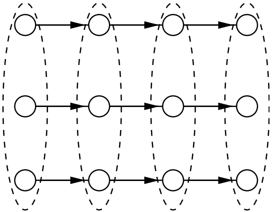
- The induced dependencies (cliques as dotted ellipses)



- Tree width of this model is? M

Structured Mean Field Factorial HMMs

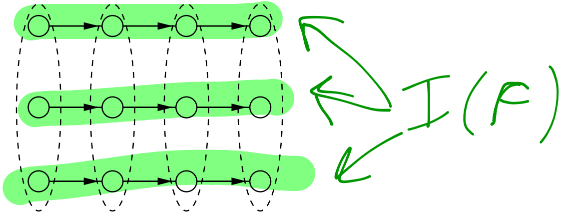
- The induced dependencies (cliques as dotted ellipses)



- Tree width of this model is? M
- Thus, if r states per chain, then exact inference complexity r^{M+1}.

Structured Mean Field Factorial HMMs

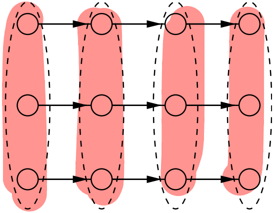
- The induced dependencies (cliques as dotted ellipses)



- 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 $O(r^2)$.

Structured Mean Field Factorial HMMs

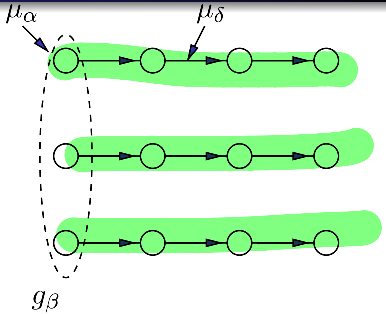
- The induced dependencies (cliques as dotted ellipses)



- 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 $O(r^M)$.

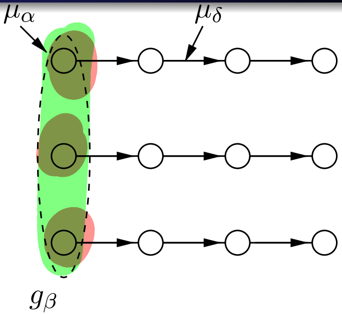
Structured Mean Field Factorial HMMs

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



Structured Mean Field Factorial HMMs

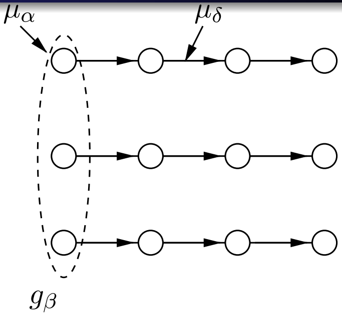
- A "natural" choice of approximating distribution is a set of coupled chains, natural, perhaps primarily for computational reasons.



- 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.

Structured Mean Field Factorial HMMs

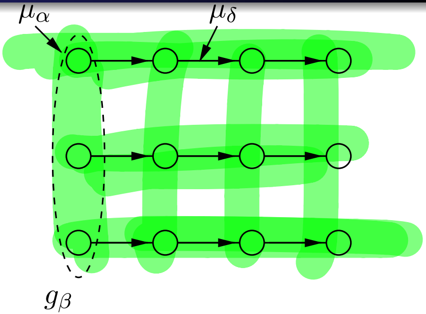
- A "natural" choice of approximating distribution is a set of coupled chains, natural, perhaps primarily for computational reasons.



- 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.

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.

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?

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.

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.

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 $\theta = (\theta_\alpha, \alpha \in \mathcal{I})$.

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.

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.

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.

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.

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.

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

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

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 $\mathcal{I}(F)$.

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 $\mu(F) \in \mathcal{M}(F) \subseteq \mathbb{R}^{d(F)}$.

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 $H(\mu(F)) \triangleq H(p_{\mu(F)}) = -A^*(\mu(F))$.

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.

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 $\mu = \mathbb{E}_{p_\theta}[\phi(X)]$.

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.

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

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

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.

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

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

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

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 $F \in \mathfrak{D}$.

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 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

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

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).

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.

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 $\mu(F)$).

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.

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:

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

$$= \bigcap_{F \in \mathfrak{D}} \mathcal{M}_F(G) \tag{18.15}$$

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:

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

$$= \bigcap_{F \in \mathfrak{D}} \mathcal{M}_F(G) \tag{18.15}$$

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

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:

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

$$= \bigcap_{F \in \mathfrak{D}} \mathcal{M}_F(G) \tag{18.15}$$

- Note this is an outer bound i.e., $\mathcal{L}(G;\mathfrak{D}) \supseteq \mathcal{M}(G)$ since $\mathcal{M}(F) \supseteq \mathcal{M}(G)$ — 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.

$$B_{\mathfrak{D}}(\theta;\rho) \triangleq \sup_{\tau \in \mathcal{L}(G;\mathfrak{D})} \left\{ \langle \tau, \theta \rangle + \sum_{F \in \mathfrak{D}} \rho(F) H(\tau(F)) \right\} \tag{18.16}$$

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.

$$B_{\mathfrak{D}}(\theta;\rho) \triangleq \sup_{\tau \in \mathcal{L}(G;\mathfrak{D})} \left\{ \langle \tau, \theta \rangle + \sum_{F \in \mathfrak{D}} \rho(F) H(\tau(F)) \right\} \tag{18.16}$$

- Objective is convex in θ since it is a max over a set of affine functions of θ (i.e., $g(\theta) = \max_\tau \langle \tau, \theta \rangle + c_\tau$)

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.

min

$$B_{\mathfrak{D}}(\theta;\rho) \triangleq \sup_{\tau \in \mathcal{L}(G;\mathfrak{D})} \left\{ \langle \tau, \theta \rangle + \sum_{F \in \mathfrak{D}} \rho(F) H(\tau(F)) \right\} \tag{18.16}$$

- 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!

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.

$$B_{\mathfrak{D}}(\theta;\rho) \triangleq \sup_{\tau \in \mathcal{L}(G;\mathfrak{D})} \left\{ \langle \tau, \theta \rangle + \sum_{F \in \mathfrak{D}} \rho(F) H(\tau(F)) \right\} \tag{18.16}$$

- 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 $\mathcal{M}(G)$

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.

$$B_{\mathfrak{D}}(\theta;\rho) \triangleq \sup_{\tau \in \mathcal{L}(G;\mathfrak{D})} \left\{ \langle \tau, \theta \rangle + \sum_{F \in \mathfrak{D}} \rho(F) H(\tau(F)) \right\} \tag{18.16}$$

- 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.

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

$$p_\theta(x) \propto \exp\left\{\sum_{s\in V}\theta_s(x_s) + \sum_{(s,t)\in E}\theta_{st}(x_s, x_t)\right\} \tag{18.17}$$

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

$$p_\theta(x) \propto \exp\left\{\sum_{s\in V}\theta_s(x_s) + \sum_{(s,t)\in E}\theta_{st}(x_s,x_t)\right\} \tag{18.17}$$

- Let $\mathfrak{T}$ be a set of all spanning trees T of G, and let ρ be a distribution over them, $\sum_{T\in\mathfrak{T}}\rho(T) = 1$.

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

$$p_\theta(x) \propto \exp\left\{\sum_{s \in V} \theta_s(x_s) + \sum_{(s,t)\in E} \theta_{st}(x_s, x_t)\right\} \tag{18.17}$$

- 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 $H(\mu) \leq \sum_{T \in \mathfrak{T}} \rho(T) H(\mu(T))$

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

$$p_\theta(x) \propto \exp\left\{ \sum_{s \in V} \theta_s(x_s) + \sum_{(s,t) \in E} \theta_{st}(x_s, x_t) \right\} \tag{18.17}$$

- 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.,

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

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

$$p_\theta(x) \propto \exp\left\{ \sum_{s \in V} \theta_s(x_s) + \sum_{(s,t) \in E} \theta_{st}(x_s, x_t) \right\} \tag{18.17}$$

- 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.,

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

- 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., $\rho_s = 1, \forall s \in V$.

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 $\sum_{s \in V} H_s(\mu_s)$.

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 $\mathfrak{T}$.

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

$$H(\mu) \leq \sum_{s \in V} H_s(\mu_s) - \sum_{(s,t) \in E} \rho_{st} I_{st}(\mu_{st}) \tag{18.19}$$

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}.

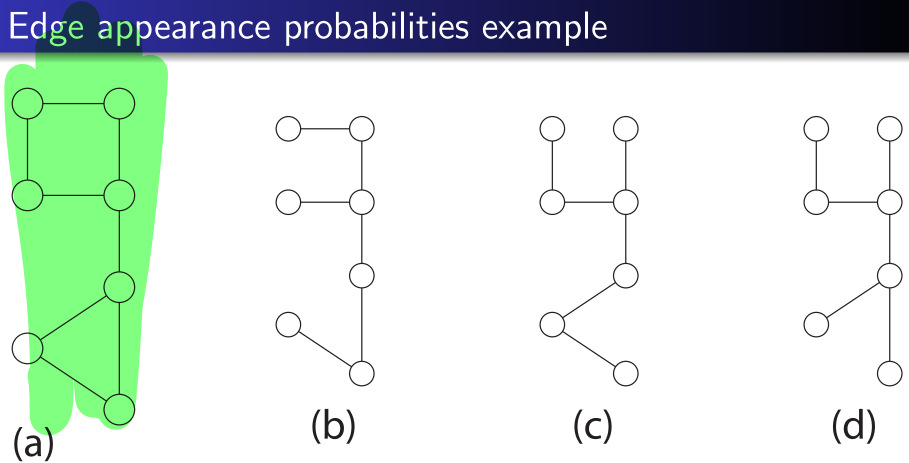
Tree-reweighted sum-product and Bethe

- 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

$$H(\mu) \leq \sum_{s \in V} H_s(\mu_s) - \sum_{(s,t) \in E} \rho_{st} I_{st}(\mu_{st}) \tag{18.19}$$

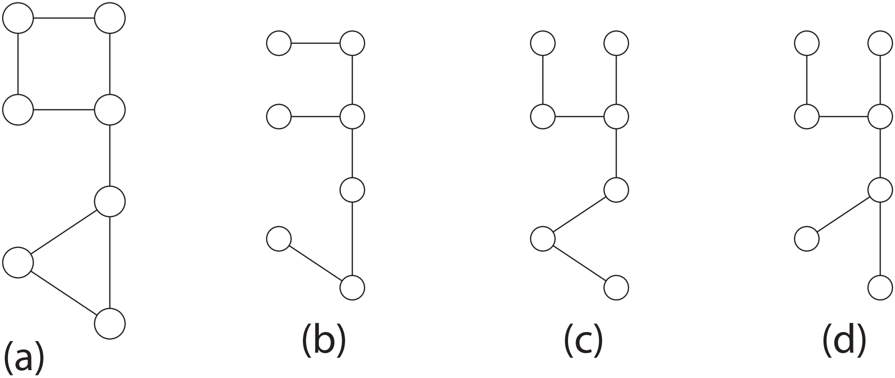
  Note right hand sum is over all E (not just a given spanning tree) and terms are weighted by probability of the given edge ρ_{st}.
- ρ_{st} is edge appearance probability, $\rho = (\rho_{st}, (s,t) \in E)$ is spanning tree polytope.

Edge appearance probabilities example



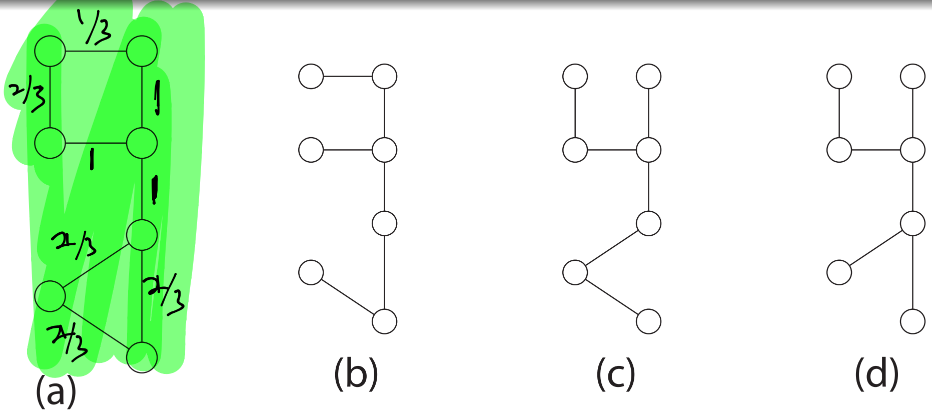
- (a) a graph $G = (V, E)$ with $m = |V| = 7$

Edge appearance probabilities example



- (a) a graph $G = (V, E)$ with $m = |V| = 7$
- (b), (c), and (d) various spanning trees, each with probability $1/3$.

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 $\mathcal{M}$.

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.

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).

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.

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.

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 $\mathbb{L}(G)$.

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):*

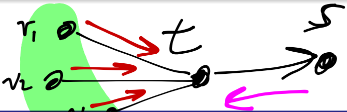
$$B_{\mathfrak{T}}(\theta;\rho) = \max_{\tau \in \mathbb{L}(G)} \left\{ \langle \tau, \theta \rangle + \sum_{s \in V} H_s(\tau_s) - \sum_{(s,t) \in E} \rho_{st} I_{st}(\tau_{st}) \right\} \tag{18.20}$$

$$\geq A(\theta) \tag{18.21}$$

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*

$$M_{t\to s}(x_s) \leftarrow \kappa \sum_{x_t'\in\mathcal{X}_t} \varphi_{st}(x_s, x_t') \frac{\prod_{v\in N(t)\setminus\{s\}} [M_{v\to t}(x_t')]^{\rho_{vt}}}{[M_{s\to t}(x_t')]^{(1-\rho_{ts})}} \tag{18.22}$$

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.

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!!

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.

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).

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

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

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

More on spanning tree polytope

- Spanning tree polytope takes the form

$$\rho \in \text{conv}(\{\mathbf{1}_T : T \in \mathfrak{T}\}) \tag{18.24}$$

where $\mathfrak{T}$ is set of all spanning trees.

More on spanning tree polytope

- Spanning tree polytope takes the form

$$\rho \in \operatorname{conv}(\{\mathbf{1}_T : T \in \mathfrak{T}\}) \tag{18.24}$$

where $\mathfrak{T}$ is set of all spanning trees.

- Consider graphic matroid on $G = (V, E)$ with rank function $r(A)$ for any $A \subseteq E$.

More on spanning tree polytope

- Spanning tree polytope takes the form

$$\rho \in \operatorname{conv}(\{\mathbf{1}_T : T \in \mathfrak{T}\}) \tag{18.24}$$

  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 $|A| = m - 1$.

More on spanning tree polytope

- Spanning tree polytope takes the form

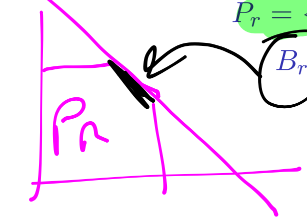
$$\rho \in \operatorname{conv}(\{\mathbf{1}_T : T \in \mathfrak{T}\}) \tag{18.24}$$

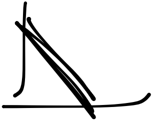
where $\mathfrak{T}$ is set of all spanning trees.

- 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:

$$P_r = \{x \in \mathbb{R}_+^E : x(A) \leq r(A), \forall A \subseteq E\} \tag{18.25}$$

$$B_r = P_r \cap \{x \in \mathbb{R}_+^E : x(E) = r(E)\} \tag{18.26}$$





More on spanning tree polytope

- Spanning tree polytope takes the form

$$\rho \in \operatorname{conv}(\{\mathbf{1}_T : T \in \mathfrak{T}\}) \tag{18.24}$$

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:

$$P_r = \left\{x \in \mathbb{R}_+^E : x(A) \le r(A), \forall A \subseteq E\right\} \tag{18.25}$$

$$B_r = P_r \cap \left\{x \in \mathbb{R}_+^E : x(E) = r(E)\right\} \tag{18.26}$$

- Then if T is a spanning tree, $\mathbf{1}_T \in B_r$, and $B_r = \operatorname{conv}(\{\mathbf{1}_T : T \in \mathfrak{T}\})$.

More on spanning tree polytope

- Spanning tree polytope takes the form

$$\rho \in \mathrm{conv}(\{\mathbf{1}_T : T \in \mathfrak{T}\}) \tag{18.24}$$

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:

$$P_r = \left\{x \in \mathbb{R}_+^E : x(A) \leq r(A), \forall A \subseteq E\right\} \tag{18.25}$$

$$B_r = P_r \cap \left\{x \in \mathbb{R}_+^E : x(E) = r(E)\right\} \tag{18.26}$$

- 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.

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 follo~~wing form:

$$\tau_s^*(x_s) = \kappa \exp\{\theta_s(x_s)\} \prod_{v \in N(s)} [M^*_{v \to s}(x_s)]^{\rho_{vs}} \tag{18.27}$$

$$\tau_{st}^*(x_s, x_t) = \kappa \varphi_{st}(x_s, x_t) \frac{\prod_{v \in N(s) \setminus t} [M^*_{vs}(x_s)]^{\rho_{vs}} \prod_{v \in N(t) \setminus s} [M^*_{vt}(x_t)]^{\rho_{vt}}}{[M^*_{ts}(x_s)]^{(1-\rho_{st})} [M^*_{st}(x_t)]^{(1-\rho_{ts})}} \tag{18.28}$$

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.

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. $\mathfrak{T}(t)$.

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

$$H(\mu) \leq E_\rho[H(\mu(T))] = \sum_{T \in \mathfrak{T}(t)} \rho(T) H(\mu(T)) \tag{18.29}$$

but here T is over all valid k-trees.

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

$$H(\mu) \leq E_\rho[H(\mu(T))] = \sum_{T \in \mathfrak{T}(t)} \rho(T) H(\mu(T)) \tag{18.29}$$

but here T is over all valid k-trees.

- This leads to a convexified Kikuchi variational problem

$$A(\theta) \leq B_{\mathfrak{B}(t)}(\theta; \rho) = \max_{\tau \in \mathbb{L}(G)} \{\langle \tau, \theta \rangle + \mathbb{E}_\rho[H(\tau(T))]\} \tag{18.30}$$

same form (but different than) before.

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

$$H(\mu) \leq E_\rho[H(\mu(T))] = \sum_{T \in \mathfrak{T}(t)} \rho(T) H(\mu(T)) \tag{18.29}$$

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

$$A(\theta) \leq B_{\mathfrak{B}(t)}(\theta; \rho) = \max_{\tau \in \mathbb{L}(G)} \{\langle \tau, \theta \rangle + \mathbb{E}_\rho[H(\tau(T))]\} \tag{18.30}$$

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

Reweighted EP

- Other variational variants have convexified version.

Reweighted EP

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

$$H_{\text{ep}}(\tau, \tilde{\tau}; \rho) = H(\tau) + \sum_{\ell=1}^{d_I} \rho(\ell)[H(\tau, \tilde{\tau}^{\ell}) - H(\tau)] \quad (18.31)$$

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

Reweighted EP

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

$$H_{\text{ep}}(\tau, \tilde{\tau}; \rho) = H(\tau) + \sum_{\ell=1}^{d_I} \rho(\ell)[H(\tau, \tilde{\tau}^{\ell}) - H(\tau)] \tag{18.31}$$

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

- In this case, reweighted entropy is concave!

Reweighted EP

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

$$H_{\text{ep}}(\tau, \tilde{\tau}; \rho) = H(\tau) + \sum_{\ell=1}^{d_I} \rho(\ell)[H(\tau, \tilde{\tau}^{\ell}) - H(\tau)] \tag{18.31}$$

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 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 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:

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 $\mathcal{M}$.

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.

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 Approximations we cover

1. Set $\mathcal{M} \leftarrow \mathbb{L}$ and $-A^*(\mu) \leftarrow H_{\text{Bethe}}(\tau)$ to get Bethe variational approximation, LBP fixed point.
2. Set $\mathcal{M} \leftarrow \mathbb{L}_t(G)$ (hypergraph marginal polytope), $-A^*(\mu) \leftarrow H_{\text{app}}(\tau)$ where $H_{\text{app}} = \sum_{g \in E} c(g) H_g(\tau_g)$ (via Möbius) to get Kikuchi variational approximation, message passing on hypergraphs.
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.
4. Mean field (from variational perspective) is (with $\mathcal{M}_F(G) \subseteq \mathcal{M}$) **l.b.**:

$$A(\theta) \geq \max_{\mu \in \mathcal{M}_F(G)} \{\langle \mu, \theta \rangle - A_F^*(\mu)\} = A_{\text{mf}}(\theta) \tag{18.1}$$

5. Upper bound Convexified/tree reweighted LBP, entropy upper bounds $H(\tau(F))$ for all members $F \in \mathfrak{D}$ of tractable substructures. Get **U.b.**:

$$A(\theta) \leq B_{\mathfrak{D}}(\theta; \rho) \triangleq \sup_{\tau \in \mathcal{L}(G; \mathfrak{D})} \left\{ \langle \tau, \theta \rangle + \sum_{F \in \mathfrak{D}} \rho(F) H(\tau(F)) \right\} \tag{18.2}$$

with $\mathcal{L}(G; \mathfrak{D}) = \bigcap_{F \in \mathfrak{D}} \mathcal{M}(F)$

MPE - most probable explanation

- In many cases, we care not to sum over x in $\sum_x p(x)$ but instead to compute $x^* \in \operatorname{argmax}_{x \in \mathsf{D}_X} p(x)$.

MPE - most probable explanation

- In many cases, we care not to sum over x in $\sum_x p(x)$ but instead to compute $x^* \in \operatorname{argmax}_{x \in \mathsf{D}_X} p(x)$.
- 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.

MPE - most probable explanation

- In many cases, we care not to sum over x in $\sum_x p(x)$ but instead to compute $x^* \in \operatorname{argmax}_{x \in \mathsf{D}_X} p(x)$.
- 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.

MPE - most probable explanation

- In many cases, we care not to sum over x in $\sum_x p(x)$ but instead to compute $x^* \in \operatorname{argmax}_{x \in \mathsf{D}_X} p(x)$.
- 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.

MPE - most probable explanation

- In many cases, we care not to sum over x in $\sum_x p(x)$ but instead to compute $x^* \in \operatorname{argmax}_{x \in \mathsf{D}_X} p(x)$.
- 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

$$\operatorname*{argmax}_{x \in \mathsf{D}_{X^m}} p(x) = \{x \in \mathsf{D}_{X^m} : p_\theta(x) \geq p_\theta(y), \forall y \in \mathsf{D}_{X^m}\} \qquad (18.32)$$

MPE - most probable explanation

- MPE again

$$\operatorname*{argmax}_{x \in \mathsf{D}_{X^m}} p(x) = \{x \in \mathsf{D}_{X^m} : p_\theta(x) \geq p_\theta(y), \forall y \in \mathsf{D}_{X^m}\} \tag{18.32}$$

- Since we are using exponential family models, we have

$$\operatorname*{argmax}_{x \in \mathsf{D}_{X^m}} p(x) = \operatorname*{argmax}_{x \in \mathsf{D}_{X^m}} \langle \theta, \phi(x) \rangle = \operatorname*{argmin}_{x \in \mathsf{D}_{X^m}} E[x] \tag{18.33}$$

  i.e., cumulant function isn't required for computation. $E[x] = -\langle \theta, \phi(x) \rangle$ is seen as an "energy" function.

MPE - most probable explanation

- MPE again

$$\operatorname*{argmax}_{x \in \mathsf{D}_{X^m}} p(x) = \{x \in \mathsf{D}_{X^m} : p_\theta(x) \geq p_\theta(y), \forall y \in \mathsf{D}_{X^m}\} \tag{18.32}$$

- Since we are using exponential family models, we have

$$\operatorname*{argmax}_{x \in \mathsf{D}_{X^m}} p(x) = \operatorname*{argmax}_{x \in \mathsf{D}_{X^m}} \langle \theta, \phi(x) \rangle = \operatorname*{argmin}_{x \in \mathsf{D}_{X^m}} E[x] \tag{18.33}$$

  i.e., cumulant function isn't required for computation. $E[x] = -\langle \theta, \phi(x) \rangle$ is seen as an "energy" function.

- But it is related. Recall cumulant function

$$A(\theta) = \log \int \exp\{\langle \theta, \phi(x) \rangle\} d\nu(x) \tag{18.34}$$

$$= \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{18.35}$$

MPE - and variational

- Considering $p_\theta(x) = \exp\left\{\langle\theta, \phi(x)\rangle - A(\theta)\right\}$.

MPE - and variational

- Considering $p_\theta(x) = \exp\{\langle\theta, \phi(x)\rangle - A(\theta)\}$.
- Let $\beta \in \mathbb{R}_+$ be a positive scalar.

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 θ 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 $\beta \to \infty$.

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 θ 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 β get large.

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 θ 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 β get large.
- Moreover, with respect to mean parameters, the maximum mean should (intuitively) fall on a vertex.

MPE - and variational

We have following theorem.

Theorem 18.6.1

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

$$\max_{x \in \mathsf{D}_{X^m}} \langle \theta, \phi(x) \rangle = \max_{\mu \in \mathcal{M}} \langle \theta, \mu \rangle, \text{ and} \tag{18.36}$$

$$\max_{x \in \mathsf{D}_{X^m}} \langle \theta, \phi(x) \rangle = \lim_{\beta \to \infty} \frac{A(\beta\theta)}{\beta} \tag{18.37}$$

- First equation shows how MPE can be seen as a LP over convex set $\mathcal{M}$.

MPE - and variational

We have following theorem.

Theorem 18.6.1

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

$$\max_{x \in \mathsf{D}_{X^m}} \langle \theta, \phi(x) \rangle = \max_{\mu \in \mathcal{M}} \langle \theta, \mu \rangle \,, \text{ and} \tag{18.36}$$

$$\max_{x \in \mathsf{D}_{X^m}} \langle \theta, \phi(x) \rangle = \lim_{\beta \to \infty} \frac{A(\beta\theta)}{\beta} \tag{18.37}$$

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

MPE - and variational

We have following theorem.

Theorem 18.6.1

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

$$\max_{x \in \mathsf{D}_{X^m}} \langle \theta, \phi(x) \rangle = \max_{\mu \in \mathcal{M}} \langle \theta, \mu \rangle, \text{ and} \tag{18.36}$$

$$\max_{x \in \mathsf{D}_{X^m}} \langle \theta, \phi(x) \rangle = \lim_{\beta \to \infty} \frac{A(\beta\theta)}{\beta} \tag{18.37}$$

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

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.

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)

$$\max_{x \in \mathsf{D}_{X^m}} \left[\sum_{s \in V} \theta_s(x_s) + \sum_{(s,t) \in E} \theta_{st}(x_s, x_t) \right] = \max_{\mu \in \mathbb{L}(T)} \langle \mu, \theta \rangle \tag{18.38}$$

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)

$$\max_{x \in \mathsf{D}_{X^m}} \left[\sum_{s \in V} \theta_s(x_s) + \sum_{(s,t) \in E} \theta_{st}(x_s, x_t) \right] = \max_{\mu \in \mathbb{L}(T)} \langle \mu, \theta \rangle \tag{18.38}$$

- Right hand side is a LP over a simple polytope, the marginal polytope for trees $\mathbb{L}(T)$.

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)

$$\max_{x \in \mathsf{D}_{X^m}} \left[\sum_{s \in V} \theta_s(x_s) + \sum_{(s,t) \in E} \theta_{st}(x_s, x_t) \right] = \max_{\mu \in \mathbb{L}(T)} \langle \mu, \theta \rangle \qquad (18.38)$$

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

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)

$$\max_{x \in \mathsf{D}_{X^m}} \left[\sum_{s \in V} \theta_s(x_s) + \sum_{(s,t) \in E} \theta_{st}(x_s, x_t) \right] = \max_{\mu \in \mathbb{L}(T)} \langle \mu, \theta \rangle \qquad (18.38)$$

- Right hand side is a LP over a simple polytope, the marginal polytope for trees $\mathbb{L}(T)$.
- 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:

$$M_{t \to s}(x_s) \leftarrow \kappa \max_{x_t \in \mathsf{D}_{X_t}} \left[\exp\{\theta_{st}(x_s, x_t) + \theta_t(x_t)\} \prod_{u \in N(t) \setminus s} M_{u \to t}(x_t) \right]$$

Max-Product and LP Duality

Theorem 18.6.2

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

$$\mathcal{Q}(\lambda) = \max_{\mu \in \mathbb{N}} \mathcal{L}(\mu; \lambda), \text{ where} \tag{18.40}$$

$$L(\mu; \lambda) = \langle \theta, \mu \rangle + \sum_{(s,t) \in E(T)} \left[\sum_{x_s} \lambda_{ts}(x_s) C_{ts}(x_s) + \sum_{x_t} \lambda_{st}(x_t) C_{st}(x_t) \right] \tag{18.41}$$

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 $\min_\lambda Q(\lambda)$.

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