EE512A – Advanced Inference in Graphical Models

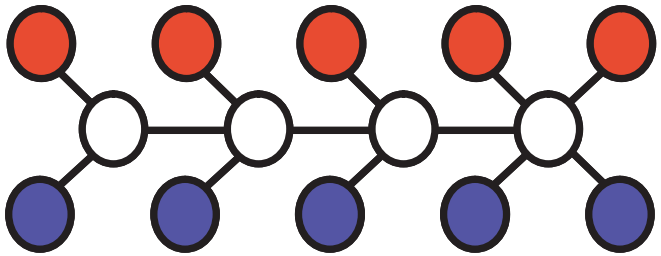
— Fall Quarter, Lecture 17 —

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

Nov 26th, 2014



Announcements

Happy Thanksgiving!! ☺

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 1,2, 3, 4 in this book. Read chapter 5.
- Also should read "Divergence measures and message passing" by Thomas Minka, and "Structured Region Graphs: Morphing EP into GBP", by Welling, Minka, and Teh.
- Assignment due Wednesday (Nov 26th) night, 11:45pm. Final project proposal updates and progress report (one page max).

Class Road Map - EE512a

- L1 (9/29): Introduction, Families, Semantics
- L2 (10/1): MRFs, elimination, Inference on Trees
- L3 (10/6): Tree inference, message passing, more general queries, non-tree)
- L4 (10/8): Non-trees, perfect elimination, triangulated graphs
- L5 (10/13): triangulated graphs, k-trees, the triangulation process/heuristics
- L6 (10/15): multiple queries, decomposable models, junction trees
- L7 (10/20): junction trees, begin intersection graphs
- L8 (10/22): intersection graphs, inference on junction trees
- L9 (10/27): inference on junction trees, semirings,
- L10 (11/3): conditioning, hardness, LBP
- L11 (11/5): LBP, exponential models,
- L12 (11/10): exponential models, mean params and polytopes,
- L13 (11/12): polytopes, tree outer bound, Bethe entropy approx.
- L14 (11/17): Bethe entropy approx, loop series correction
- L15 (11/19): Hypergraphs, posets, Mobius, Kikuchi
- L16 (11/24): Kikuchi, Expectation Propagation
- L17 (11/26): Expectation Propagation, Mean Field
- L18 (12/1):
- L19 (12/3):
- Final Presentations: (12/10):

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

Term Decoupling in EP

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

$$\phi \triangleq (\phi_1, \phi_2, \ldots, \phi_{d_T}) \tag{17.5}$$

- Intractable component

$$\Phi \triangleq (\Phi^1, \Phi^2, \ldots, \Phi^{d_I}) \tag{17.6}$$

- ϕ_i are typically univariate, while Φ^i are typically multivariate (b-dimensional we'll assume), although this need not always be the case (but will be for our exposition).
- Consider exponential families associated with subcollection (ϕ, Φ).

Associated Distributions: base and i-augmented

- The associated exponential family

$$p(x;\theta,\tilde{\theta}) \propto \exp\left(\langle\theta,\phi(x)\rangle\right)\exp\left(\left\langle\tilde{\theta},\Phi(x)\right\rangle\right) \tag{17.7}$$

$$= \exp\left(\langle\theta,\phi(x)\rangle\right)\prod_{i=1}^{d_I}\exp\left(\left\langle\tilde{\theta}^i,\Phi^i(x)\right\rangle\right) \tag{17.8}$$

- Base model is tractable

$$p(x;\theta,\vec{0}) \propto \exp\left(\langle\theta,\phi(x)\rangle\right) \tag{17.9}$$

- Φ^i-augmented model

$$p(x;\theta,\tilde{\theta}^i) \propto \exp\left(\langle\theta,\phi(x)\rangle\right)\exp\left(\left\langle\tilde{\theta}^i,\Phi^i(x)\right\rangle\right) \tag{17.10}$$

New EP-based outer bound

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

$$\Pi^i(\tau, \tilde{\tau}) \to (\tau, \tilde{\tau}^i) \tag{17.14}$$

  This operator simply removes all but $\tilde{\tau}^i$ from $\tilde{\tau}$.
- Define outer bound on true means $\mathcal{M}(\phi, \Phi)$ (which is still convex)

$$\mathcal{L}(\phi, \Phi) = \left\{(\tau, \tilde{\tau}) | \tau \in \mathcal{M}(\phi), \Pi^i(\tau, \tilde{\tau}) \in \mathcal{M}(\phi, \Phi^i),\ \forall i\right\} \tag{17.15}$$

- Note, based on a set of projections onto $\mathcal{M}(\phi, \Phi^i)$.
- Outer bound, i.e., $\mathcal{M}(\phi, \Phi) \subseteq \mathcal{L}(\phi, \Phi)$, since:

$$\tau \in \mathcal{M}(\phi) \Leftrightarrow \exists p \text{ s.t. } \tau = E_p[\phi(X)] \tag{17.16}$$

$$(\tau, \tilde{\tau}) \in \mathcal{L}(\phi, \Phi) \Leftrightarrow \tau \in \mathcal{M}(\phi) \ \&\ \exists p \text{ s.t. } (\tau, \tilde{\tau}^i) = E_p[\phi(X), \Phi^i(X)] \tag{17.17}$$

$$(\tau, \tilde{\tau}) \in \mathcal{M}(\phi, \Phi) \Leftrightarrow \exists p \text{ s.t. } (\tau, \tilde{\tau}) = E_p[\phi(X), \Phi(X)] \tag{17.18}$$

- If Φ^i are edges of a graph (i.e. local consistency) then we get standard $\mathbb{L}$ outer bound we saw before with Bethe approximation

EP outer bound entropy and opt

- For any mean parms $(\tau, \tilde{\tau}) \in \mathcal{L}(\phi, \Phi)$: A) There is a member of the ϕ-exponential family which mean parameters τ with entropy $H(\tau)$; B) Also, for $i = 1 \ldots d_I$, there is a member of the (ϕ, Φ^i)-exponential family with mean parameters $(\tau, \tilde{\tau}^i)$ with entropy $H(\tau, \tilde{\tau}^i)$.
- Both entropy forms are easy to compute, and so is a new entropy approximation:

$$H(\tau, \tilde{\tau}) \approx H_{\mathsf{ep}}(\tau, \tilde{\tau}) \triangleq H(\tau) + \sum_{\ell=1}^{d_I} \left[H(\tau, \tilde{\tau}^l) - H(\tau) \right] \tag{17.14}$$

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

$$\max_{(\tau, \tilde{\tau}) \in \mathcal{L}(\phi, \Phi)} \left\{ \langle \tau, \theta \rangle + \left\langle \tilde{\tau}, \tilde{\theta} \right\rangle + H_{\mathsf{ep}}(\tau, \tilde{\tau}) \right\} \tag{17.15}$$

- This characterizes the EP algorithms.
- Given graph $G = (V, E)$ when we take ϕ to be unaries V and Φ to be edges E, we exactly recover Bethe approximation.

Lagrangian optimization setup

- Make d_I duplicates of vector $\tau \in \mathbb{R}^{d_T}$, call them $\eta^i \in \mathbb{R}^{d_T}$ for $i \in [d_T]$.
- This gives large set of pseudo-mean parameters

$$\left\{\tau, (\eta^i, \tilde{\tau}^i), i \in [d_I]\right\} \in \mathbb{R}^{d_T} \times (\mathbb{R}^{d_T} \times \mathbb{R}^b)^{d_I} \tag{17.14}$$

- We arrive at the optimization:

$$\max_{\left\{\tau, \{(\eta^i, \tilde{\tau}^i)\}_i\right\}} \left\{ \langle \tau, \theta \rangle + \sum_{i=1}^{d_I} \left\langle \tilde{\tau}^i, \tilde{\theta}^i \right\rangle + H(\tau) + \sum_{i=1}^{d_I} \left[H(\eta^i, \tilde{\tau}^i) - H(\eta^i) \right] \right\} \tag{17.15}$$

  subject to $\tau \in \mathcal{M}(\phi)$, and for all i that $\tau = \eta^i$ and that $(\eta^i, \tilde{\tau}^i) \in \mathcal{M}(\phi, \Phi^i)$.
- Use Lagrange multipliers to impose constraint $\eta^i = \tau$ for all i, and for the rest of the constraints too.

Moment Matching → Expectation Propagation Updates

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

$$q^i(x; \theta, \tilde{\theta}^i, \lambda) \propto \exp\left(\left\langle \theta + \sum_{\ell \neq i} \lambda^l, \phi(x) \right\rangle + \left\langle \tilde{\theta}^i, \Phi^i(x) \right\rangle\right), \quad (17.19)$$

compute the mean parameters η^i as follows:

$$\eta^{i(n)} = \int q^{i(n)}(x)\phi(x)\nu(dx) = \mathbb{E}_{q^{i(n)}}[\phi(X)] \quad (17.20)$$

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

$$\mathbb{E}_q[\phi(X)] = \eta^{i(n)} \quad (17.21)$$

5. This is a KL-divergence minimization step, but done w. exponential family models which thus corresponds to moment-matching.

Variational Approach Amenable to Approximation

- Original variational representation of log partition function

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{17.1}$$

where dual takes form:

$$A^*(\mu) = \sup_{\theta \in \Omega} (\langle \theta, \mu \rangle - A(\theta)) = \begin{cases} -H(p_{\theta(\mu)}) & \text{if } \mu \in \mathcal{M}^\circ \\ +\infty & \text{if } \mu \notin \overline{\mathcal{M}} \end{cases} \tag{17.2}$$

Variational Approach Amenable to Approximation

- Original variational representation of log partition function

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{17.1}$$

where dual takes form:

$$A^*(\mu) = \sup_{\theta \in \Omega} (\langle \theta, \mu \rangle - A(\theta)) = \begin{cases} -H(p_{\theta(\mu)}) & \text{if } \mu \in \mathcal{M}^\circ \\ +\infty & \text{if } \mu \notin \overline{\mathcal{M}} \end{cases} \tag{17.2}$$

- Given efficient expression for $A(\theta)$, we can compute marginals of interest.

Variational Approach Amenable to Approximation

- Original variational representation of log partition function

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{17.1}$$

where dual takes form:

$$A^*(\mu) = \sup_{\theta \in \Omega} (\langle \theta, \mu \rangle - A(\theta)) = \begin{cases} -H(p_{\theta(\mu)}) & \text{if } \mu \in \mathcal{M}^\circ \\ +\infty & \text{if } \mu \notin \overline{\mathcal{M}} \end{cases} \tag{17.2}$$

- 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.

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.

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.

Example: Sum-Product, Bethe, and EP: distributions

- EP generalizes sum-product and Bethe approximation we saw from a few lectures ago.

Example: Sum-Product, Bethe, and EP: distributions

- EP generalizes sum-product and Bethe approximation we saw from a few lectures ago.
- Recall, general graph $G = (V, E)$ and we have parameters and statistics associated with each node $\phi_s(x_s)$ for $s \in V$ and each edge $\phi_{u,v}(x_u, x_v)$ for $(u, v) \in E(G)$.

Example: Sum-Product, Bethe, and EP: distributions

- EP generalizes sum-product and Bethe approximation we saw from a few lectures ago.
- Recall, general graph $G = (V, E)$ and we have parameters and statistics associated with each node $\phi_s(x_s)$ for $s \in V$ and each edge $\phi_{u,v}(x_u, x_v)$ for $(u, v) \in E(G)$.
- Base distribution is only the nodes (fully factored independent distribuiton)

$$p(x; \phi_1, \ldots, \phi_m, \vec{0}) \propto \prod_{v \in V} \exp(\theta_s(x_s)) \tag{17.1}$$

Example: Sum-Product, Bethe, and EP: distributions

- EP generalizes sum-product and Bethe approximation we saw from a few lectures ago.
- Recall, general graph $G = (V, E)$ and we have parameters and statistics associated with each node $\phi_s(x_s)$ for $s \in V$ and each edge $\phi_{u,v}(x_u, x_v)$ for $(u, v) \in E(G)$.
- Base distribution is only the nodes (fully factored independent distribuiton)

$$p(x; \phi_1, \ldots, \phi_m, \vec{0}) \propto \prod_{v \in V} \exp(\theta_s(x_s)) \tag{17.1}$$

- Each Φ^i corresponds to an edge (e.g., $i = (u, v)$ for some edge $(u, v) \in E(G)$). Hence, $\Phi^{u,v}$-augmented distribution takes form:

$$p(x; \phi_1, \ldots, \phi_m, \phi_{uv}) \propto \prod_{v \in V} \exp(\theta_s(x_s)) \exp(\theta_{uv}(x_u, x_v)) \tag{17.2}$$

Example: Sum-Product, Bethe, and EP: entropies

- Base entropy is sum of node marginal entropies

$$H(\tau_1, \ldots, \tau_m) = \sum_{s \in V} H(\tau_s) \tag{17.3}$$

Example: Sum-Product, Bethe, and EP: entropies

- Base entropy is sum of node marginal entropies

$$H(\tau_1, \ldots, \tau_m) = \sum_{s \in V} H(\tau_s) \tag{17.3}$$

- Augmented entropy takes the form

$$H(\tau_1, \ldots, \tau_m, \tau_{uv}) = \sum_{s \in V \setminus \{u,v\}} H(\tau_s) + H(\tau_{uv}) \tag{17.4}$$

$$= \sum_{s \in V} H(\tau_s) + [H(\tau_{uv}) - H(\tau_u) - H(\tau_v)] \tag{17.5}$$

$$= \sum_{s \in V} H(\tau_s) + I(\tau_{u,v}) \tag{17.6}$$

where $I(\tau_{u,v})$ is the mutual information between X_u and X_v under joint distribution τ_{uv}.

Example: Sum-Product, Bethe, and EP: entropies

- Base entropy is sum of node marginal entropies

$$H(\tau_1, \ldots, \tau_m) = \sum_{s \in V} H(\tau_s) \tag{17.3}$$

- Augmented entropy takes the form

$$H(\tau_1, \ldots, \tau_m, \tau_{uv}) = \sum_{s \in V \setminus \{u,v\}} H(\tau_s) + H(\tau_{uv}) \tag{17.4}$$

$$= \sum_{s \in V} H(\tau_s) + [H(\tau_{uv}) - H(\tau_u) - H(\tau_v)] \tag{17.5}$$

$$= \sum_{s \in V} H(\tau_s) + I(\tau_{u,v}) \tag{17.6}$$

where $I(\tau_{u,v})$ is the mutual information between X_u and X_v under joint distribution τ_{uv}.

- Overall EP entropy, suming over all augmentations $(u, v) \in E(G)$, is:

$$H_{\text{ep}}(\tau) = \sum_{s \in V} H(\tau_s) - \sum_{(u,v) \in E(G)} I(\tau_{uv}) \tag{17.7}$$

Example: Sum-Product, Bethe, and EP: $\mathcal{L}(\phi, \Phi)$

- the base mean parameter $\mathcal{M}(\phi)$ just asks that $\tau = (\tau_s, s \in V)$ are valid unary marginals (i.e., non-negative and sum to one, in the form of $\forall s \in V$, $0 \leq \tau_s(x_s) \leq 1$ and $\sum_{x_s} \tau_s(x_s) = 1$.

Example: Sum-Product, Bethe, and EP: $\mathcal{L}(\phi, \Phi)$

- the base mean parameter $\mathcal{M}(\phi)$ just asks that $\tau = (\tau_s, s \in V)$ are valid unary marginals (i.e., non-negative and sum to one, in the form of $\forall s \in V$, $0 \leq \tau_s(x_s) \leq 1$ and $\sum_{x_s} \tau_s(x_s) = 1$.
- Each augmentation $\mathcal{M}(\phi, \Phi^{uv})$ for edge $(u, v) \in E(G)$ also asks that τ_{uv} marginalizes down to τ_u and τ_v, i.e., $\sum_{x_v} \tau_{uv}(x_v, x_u) = \tau_u(x_u)$ and $\sum_{x_u} \tau_{uv}(x_v, x_u) = \tau_v(x_v)$.

Example: Sum-Product, Bethe, and EP: $\mathcal{L}(\phi, \Phi)$

- the base mean parameter $\mathcal{M}(\phi)$ just asks that $\tau = (\tau_s, s \in V)$ are valid unary marginals (i.e., non-negative and sum to one, in the form of $\forall s \in V$, $0 \leq \tau_s(x_s) \leq 1$ and $\sum_{x_s} \tau_s(x_s) = 1$.
- Each augmentation $\mathcal{M}(\phi, \Phi^{uv})$ for edge $(u, v) \in E(G)$ also asks that τ_{uv} marginalizes down to τ_u and τ_v, i.e., $\sum_{x_v} \tau_{uv}(x_v, x_u) = \tau_u(x_u)$ and $\sum_{x_u} \tau_{uv}(x_v, x_u) = \tau_v(x_v)$.
- Then considering $\mathcal{L}(\phi, \Phi)$ as defined, we must have for all $(u, v) \in E(G)$, $\Pi^{uv}(\tau, \tilde{\tau}) \in \mathcal{M}(\phi, \Phi^{uv})$ — this requires local consistency along all edges of the graph.

Example: Sum-Product, Bethe, and EP: $\mathcal{L}(\phi, \Phi)$

- the base mean parameter $\mathcal{M}(\phi)$ just asks that $\tau = (\tau_s, s \in V)$ are valid unary marginals (i.e., non-negative and sum to one, in the form of $\forall s \in V$, $0 \leq \tau_s(x_s) \leq 1$ and $\sum_{x_s} \tau_s(x_s) = 1$.
- Each augmentation $\mathcal{M}(\phi, \Phi^{uv})$ for edge $(u, v) \in E(G)$ also asks that τ_{uv} marginalizes down to τ_u and τ_v, i.e., $\sum_{x_v} \tau_{uv}(x_v, x_u) = \tau_u(x_u)$ and $\sum_{x_u} \tau_{uv}(x_v, x_u) = \tau_v(x_v)$.
- Then considering $\mathcal{L}(\phi, \Phi)$ as defined, we must have for all $(u, v) \in E(G)$, $\Pi^{uv}(\tau, \tilde{\tau}) \in \mathcal{M}(\phi, \Phi^{uv})$ — this requires local consistency along all edges of the graph.
- Therefore, in this case, $\mathcal{L}(\phi, \Phi)$ is the same as the local consistency (or tree-based) polytope outer bound we encountered with LBP and the Bethe approximation.

Ex: Sum-Prod., Bethe, and EP: moment matching, nodes

- The base distribution with the Lagrange multipliers has the form:

$$q(x;\theta,\lambda) \propto \prod_{s\in V} \exp(\theta_s(x_s)) \prod_{(u,v)\in E} \exp(\lambda_{uv}(x_v) + \lambda_{vu}(x_u)) \quad (17.8)$$

$$= \prod_{s\in V} \exp(\theta_s(x_s) + \sum_{t\in N(s)} \lambda_{ts}(x_s)) \quad (17.9)$$

$$\propto \prod_{s\in V} \tau_s(x_s) \quad (17.10)$$

where $\tau_s(x_s) = \exp\left(\theta_s(x_s) + \sum_{t\in N(s)} \lambda_{ts}(x_s)\right)$.

Ex: Sum-Prod., Bethe, and EP: moment matching, nodes

- The base distribution with the Lagrange multipliers has the form:

$$q(x;\theta,\lambda) \propto \prod_{s\in V} \exp(\theta_s(x_s)) \prod_{(u,v)\in E} \exp(\lambda_{uv}(x_v) + \lambda_{vu}(x_u)) \quad (17.8)$$

$$= \prod_{s\in V} \exp(\theta_s(x_s) + \sum_{t\in N(s)} \lambda_{ts}(x_s)) \quad (17.9)$$

$$\propto \prod_{s\in V} \tau_s(x_s) \quad (17.10)$$

where $\tau_s(x_s) = \exp\left(\theta_s(x_s) + \sum_{t\in N(s)} \lambda_{ts}(x_s)\right)$.

- This marginal takes the form of messages being sent along s's neighbors to node s, just like in BP.

Example: Sum-Product, Bethe, and EP: moment matching

- Augmented distribution takes the form, for edge $\ell = (u, v)$,

$$q^{(u,v)}(x;\theta,\lambda) \propto q(x;\theta,\lambda)\exp(\theta_{uv}(x_u,x_v) - \lambda_{uv}(x_v) - \lambda_{uv}(x_u))$$
$$= \left[\prod_{s\in V}\tau_s(x_s)\right]\exp(\theta_{uv}(x_u,x_v) - \lambda_{uv}(x_v) - \lambda_{uv}(x_u)) \tag{17.11}$$

Example: Sum-Product, Bethe, and EP: moment matching

- Augmented distribution takes the form, for edge $\ell = (u, v)$,

$$q^{(u,v)}(x;\theta,\lambda) \propto q(x;\theta,\lambda)\exp(\theta_{uv}(x_u,x_v) - \lambda_{uv}(x_v) - \lambda_{uv}(x_u))$$
$$= \left[\prod_{s\in V}\tau_s(x_s)\right]\exp(\theta_{uv}(x_u,x_v) - \lambda_{uv}(x_v) - \lambda_{uv}(x_u)) \tag{17.11}$$

- Then the EP algorithm (with this set of base and augmented statistics) is such that we repeated choose an edge $(u, v) \in E(G)$, form distribution above, and adjust $\lambda_{uv}(x_v)$ and $\lambda_{vu}(x_u)$ in Equation (17.8) so that the marginal distributions $\tau_v(x_v)$ and $\tau_u(x_u)$ match the marginals of the joint along this edge.

Example: Sum-Product, Bethe, and EP: moment matching

- Augmented distribution takes the form, for edge $\ell = (u, v)$,

$$q^{(u,v)}(x;\theta,\lambda) \propto q(x;\theta,\lambda)\exp(\theta_{uv}(x_u,x_v) - \lambda_{uv}(x_v) - \lambda_{uv}(x_u))$$
$$= \left[\prod_{s\in V}\tau_s(x_s)\right]\exp(\theta_{uv}(x_u,x_v) - \lambda_{uv}(x_v) - \lambda_{uv}(x_u)) \tag{17.11}$$

- Then the EP algorithm (with this set of base and augmented statistics) is such that we repeated choose an edge $(u, v) \in E(G)$, form distribution above, and adjust $\lambda_{uv}(x_v)$ and $\lambda_{vu}(x_u)$ in Equation (17.8) so that the marginal distributions $\tau_v(x_v)$ and $\tau_u(x_u)$ match the marginals of the joint along this edge.
- Key point: This marginal matching in fact correspond to the marginal updates of the standard BP algorithm!

Example: Tree-structured EP

- EP is much more general than this. In above case, base distribution was all singletons (all independent) and augmentation was edges.

Example: Tree-structured EP

- EP is much more general than this. In above case, base distribution was all singletons (all independent) and augmentation was edges.
- When base distribution is a tree, we get tree-structured EP

Example: Tree-structured EP

- EP is much more general than this. In above case, base distribution was all singletons (all independent) and augmentation was edges.
- When base distribution is a tree, we get tree-structured EP
- Start with a graph $G = (V, E)$ and form a spanning tree $T = (V, E(T))$ in any arbitrary way.

Example: Tree-structured EP

- EP is much more general than this. In above case, base distribution was all singletons (all independent) and augmentation was edges.
- When base distribution is a tree, we get tree-structured EP
- Start with a graph $G = (V, E)$ and form a spanning tree $T = (V, E(T))$ in any arbitrary way.
- Form base tree distribution as follows:

$$p(x; \theta, \vec{0}) \propto \prod_{s \in V} \exp(\theta_s(x_s)) \prod_{(s,t) \in E(T)} \exp(\theta_{st}(x_s, x_t)) \tag{17.12}$$

Example: Tree-structured EP

- EP is much more general than this. In above case, base distribution was all singletons (all independent) and augmentation was edges.
- When base distribution is a tree, we get tree-structured EP
- Start with a graph $G = (V, E)$ and form a spanning tree $T = (V, E(T))$ in any arbitrary way.
- Form base tree distribution as follows:

$$p(x; \theta, \vec{0}) \propto \prod_{s \in V} \exp(\theta_s(x_s)) \prod_{(s,t) \in E(T)} \exp(\theta_{st}(x_s, x_t)) \tag{17.12}$$

- Then, each Φ^i corresponds to an edge in $E \setminus E(T)$, and gives us, for each edge $(u, v) \in E \setminus E(T)$, the $\phi^{(u,v)}$-augmented distribution

$$p(x; \theta, \theta_{u,v}) \propto (x; \theta, \vec{0}) \exp(\theta_{u,v}(x_u, x_v)) \tag{17.13}$$

EP as variational: Summary of key points

- Fixed points of EP exist assuming Lagrangian form has at least one optimum.

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 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.

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.

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.

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).

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.

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/

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.

Mean Field

- So far, we have been using an outer bound on $\mathcal{M}$.

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.

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.

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.

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.

Tractable Families

- 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 $E_F \subseteq E$.

Tractable Families

- 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: $F = (V, \emptyset)$.

Tractable Families

- 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.

Tractable Families

- 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.

Tractable Families

- 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.

Tractable Families

- 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. The parameters that respect F are of the form:

$$\mathbb{R}^{|\mathcal{I}|} \ni \Omega(F) \triangleq \{\theta \in \Omega | \theta_\alpha = 0 \ \ \forall \alpha \in \mathcal{I} \setminus \mathcal{I}(F)\} \subseteq \Omega \qquad (17.14)$$

Tractable Families

- 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. The parameters that respect F are of the form:

$$\mathbb{R}^{|\mathcal{I}|} \ni \Omega(F) \triangleq \{\theta \in \Omega | \theta_\alpha = 0 \;\; \forall \alpha \in \mathcal{I} \setminus \mathcal{I}(F)\} \subseteq \Omega \qquad (17.14)$$

notice, all parameters associated with sufficient statistic not in $\mathcal{I}(F)$ are set to zero, those statistics are nonexistent in F.

Tractable Families

- 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. The parameters that respect F are of the form:

$$\mathbb{R}^{|\mathcal{I}|} \ni \Omega(F) \triangleq \{\theta \in \Omega | \theta_\alpha = 0 \;\; \forall \alpha \in \mathcal{I} \setminus \mathcal{I}(F)\} \subseteq \Omega \qquad (17.14)$$

  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.

Tractable Subgraphs: All Independent Example

- Ex: MRF with potential functions for nodes and edges.

Tractable Subgraphs: All Independent Example

- Ex: MRF with potential functions for nodes and edges.
- For each $(s,t) \in E(G)$, we have $\theta_{(s,t)}$.

Tractable Subgraphs: All Independent Example

- Ex: MRF with potential functions for nodes and edges.
- For each $(s,t) \in E(G)$, we have $\theta_{(s,t)}$.
- $F_0 = (V, \emptyset)$ which yields

$$\Omega(F_0) = \left\{\theta \in \Omega | \theta_{(s,t)} = 0 \;\; \forall (s,t) \in E(G)\right\} \tag{17.15}$$

Tractable Subgraphs: All Independent Example

- Ex: MRF with potential functions for nodes and edges.
- For each $(s,t) \in E(G)$, we have $\theta_{(s,t)}$.
- $F_0 = (V, \emptyset)$ which yields

$$\Omega(F_0) = \left\{\theta \in \Omega | \theta_{(s,t)} = 0 \;\; \forall (s,t) \in E(G)\right\} \tag{17.15}$$

- This is the all independence model, giving family of distributions

$$p_\theta(x) = \prod_{s \in V} p(x_s; \theta_s) \tag{17.16}$$

Tractable Subgraphs: Tree Example

- Ex: MRF with potential functions for nodes and edges.

Tractable Subgraphs: Tree Example

- Ex: MRF with potential functions for nodes and edges.
- For each $(s,t) \in E(G)$, we have $\theta_{(s,t)}$.

Tractable Subgraphs: Tree Example

- Ex: MRF with potential functions for nodes and edges.
- For each $(s,t) \in E(G)$, we have $\theta_{(s,t)}$.
- $F_T = (V,T)$ where $T \subset E$ are edges that constitute a spanning tree of G, giving

$$\Omega(F_0) = \left\{\theta \in \Omega | \theta_{(s,t)} = 0 \;\; \forall (s,t) \notin T\right\} \tag{17.17}$$

Tractable Subgraphs: Tree Example

- Ex: MRF with potential functions for nodes and edges.
- For each $(s,t) \in E(G)$, we have $\theta_{(s,t)}$.
- $F_T = (V,T)$ where $T \subset E$ are edges that constitute a spanning tree of G, giving

$$\Omega(F_0) = \left\{\theta \in \Omega | \theta_{(s,t)} = 0 \;\; \forall (s,t) \notin T\right\} \tag{17.17}$$

- This gives a tree-dependent family

$$p_\theta(x) = \prod_{s \in V} p(x_s;\theta_s) \prod_{(s,t)\in T} \frac{p(x_s,x_t;\theta_{st})}{p(x_s;\theta_s)p(x_t;\theta_t)} \tag{17.18}$$

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.

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.,

$$\mathcal{M}_F(G;\phi) = \left\{\mu \in \mathbb{R}^d | \mu = \mathbb{E}_\theta[\phi(x)] \text{ for some } \theta \in \Omega(F)\right\} \quad (17.19)$$

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.,

$$\mathcal{M}_F(G;\phi) = \left\{\mu \in \mathbb{R}^d | \mu = \mathbb{E}_\theta[\phi(x)] \text{ for some } \theta \in \Omega(F)\right\} \quad (17.19)$$

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

$$\mathcal{M}_F^\circ(G;\phi) \subseteq \mathcal{M}^\circ(G;\phi) \quad (17.20)$$

and so $\mathcal{M}_F^\circ(G;\phi)$ is an inner approximation of the set of realizable mean parameters.

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.,

$$\mathcal{M}_F(G;\phi) = \left\{\mu \in \mathbb{R}^d | \mu = \mathbb{E}_\theta[\phi(x)] \text{ for some } \theta \in \Omega(F)\right\} \quad (17.19)$$

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

$$\mathcal{M}_F^\circ(G;\phi) \subseteq \mathcal{M}^\circ(G;\phi) \quad (17.20)$$

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

Mean field variational lower bound

- Mean field methods generate lower bounds on their estimated $A(\theta)$ and approximate mean parameters $\mu = \mathbb{E}_\theta[\phi(X)]$.

Proposition 17.4.1 (mean field lower bound)

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

$$A(\theta) \geq \langle \theta, \mu \rangle - A^*(\mu) \tag{17.21}$$

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

Mean field variational lower bound

Proof.

- On the one hand, obvious due to $A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\}$

Mean field variational lower bound

Proof.

- On the one hand, obvious due to $A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\}$
- More traditional proof, let q be <u>any</u> distribution that satisfies moment matching $\mathbb{E}_q[\phi(X)] = \mu$, then:

$$A(\theta) = \log \int_{\mathcal{X}^m} \exp \langle \theta, \phi(x) \rangle \nu(dx) \tag{17.22}$$

$$= \log \int_{\mathcal{X}^m} q(x) \frac{\exp \langle \theta, \phi(x) \rangle}{q(x)} \nu(dx) \tag{17.23}$$

$$\geq \int_{\mathcal{X}^m} q(x) [\langle \theta, \phi(x) \rangle - \log q(x)] \nu(dx) \tag{17.24}$$

$$= \langle \theta, E_q[\phi(X)] \rangle - H(q) = \langle \theta, \mu \rangle - H(q) \tag{17.25}$$

Mean field variational lower bound

Proof.

- On the one hand, obvious due to $A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\}$
- More traditional proof, let q be <u>any</u> distribution that satisfies moment matching $\mathbb{E}_q[\phi(X)] = \mu$, then:

$$A(\theta) = \log \int_{\mathcal{X}^m} \exp \langle \theta, \phi(x) \rangle \nu(dx) \tag{17.22}$$

$$= \log \int_{\mathcal{X}^m} q(x) \frac{\exp \langle \theta, \phi(x) \rangle}{q(x)} \nu(dx) \tag{17.23}$$

$$\geq \int_{\mathcal{X}^m} q(x) [\langle \theta, \phi(x) \rangle - \log q(x)] \nu(dx) \tag{17.24}$$

$$= \langle \theta, E_q[\phi(X)] \rangle - H(q) = \langle \theta, \mu \rangle - H(q) \tag{17.25}$$

- If we optimize q over all $\mathcal{M}(G)$, then we'll get equality.

Mean field variational lower bound

Proof.

- On the one hand, obvious due to $A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\}$
- More traditional proof, let q be <u>any</u> distribution that satisfies moment matching $\mathbb{E}_q[\phi(X)] = \mu$, then:

$$A(\theta) = \log \int_{\mathcal{X}^m} \exp \langle \theta, \phi(x) \rangle \nu(dx) \tag{17.22}$$

$$= \log \int_{\mathcal{X}^m} q(x) \frac{\exp \langle \theta, \phi(x) \rangle}{q(x)} \nu(dx) \tag{17.23}$$

$$\geq \int_{\mathcal{X}^m} q(x) [\langle \theta, \phi(x) \rangle - \log q(x)] \nu(dx) \tag{17.24}$$

$$= \langle \theta, E_q[\phi(X)] \rangle - H(q) = \langle \theta, \mu \rangle - H(q) \tag{17.25}$$

- If we optimize q over all $\mathcal{M}(G)$, then we'll get equality.
- If we optimize q over a subset of $\mathcal{M}(G)$ (e.g., such as $\mathcal{M}_F(G)$, then we'll get inequality.

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.

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:

$$A(\theta) \geq \max_{\mu \in \mathcal{M}_F(G)} \{\langle \mu, \theta \rangle - A_F^*(\mu)\} \triangleq A_{\mathsf{MF}}(\theta) \tag{17.26}$$

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 $A^*(\mu)$.

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:

$$A(\theta) \geq \max_{\mu \in \mathcal{M}_F(G)} \{\langle \mu, \theta \rangle - A_F^*(\mu)\} \triangleq A_{\mathsf{MF}}(\theta) \tag{17.26}$$

  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.

Recall

Recall the following slide from lecture 13.

Conjugate Duality, Maximum Likelihood, Negative Entropy

Theorem 17.4.3 (Relationship between A and A^*)

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

$$A^*(\mu) = \sup_{\theta \in \Omega} (\langle \theta, \mu \rangle - A(\theta)) = \begin{cases} -H(p_{\theta(\mu)}) & \text{if } \mu \in \mathcal{M}^\circ \\ +\infty & \text{if } \mu \notin \overline{\mathcal{M}} \end{cases} \tag{17.3}$$

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

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{17.4}$$

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

$$\mu = \int_{\mathsf{D}_X} \phi(x) p_\theta(x) \nu(dx) = \mathbb{E}_\theta[\phi(X)] = \nabla A(\theta) \tag{17.5}$$

Mean field, KL-Divergence, Exponential Model Families

- The conjugae dual optimizations associated with the above, in the mean field framework has a nice interpretation in terms of minimizing a KL divergence.

Mean field, KL-Divergence, Exponential Model Families

- The conjugae dual optimizations associated with the above, in the mean field framework has a nice interpretation in terms of minimizing a KL divergence.
- In particular, mean field can be seen as finding the best, in a KL-divergence minimization sense, approximation to a distribution from among a family of tractable distributions.

Mean field, KL-Divergence, Exponential Model Families

- Given two distributions p, q, KL-Divergence of p w.r.t. q is defined as

$$D(q||p) = \int_{\mathcal{X}^m} q(x) \left[\log \frac{q(x)}{p(x)} \right] \nu(dx) \tag{17.27}$$

Mean field, KL-Divergence, Exponential Model Families

- Given two distributions p, q, KL-Divergence of p w.r.t. q is defined as

$$D(q||p) = \int_{\mathcal{X}^m} q(x) \left[\log \frac{q(x)}{p(x)} \right] \nu(dx) \tag{17.27}$$

- In summation form, we have

$$D(q||p) = \sum_{x \in \mathcal{X}^m} q(x) \left[\log \frac{q(x)}{p(x)} \right] \tag{17.28}$$

Mean field, KL-Divergence, Exponential Model Families

- Given two distributions p, q, KL-Divergence of p w.r.t. q is defined as

$$D(q||p) = \int_{\mathcal{X}^m} q(x) \left[\log \frac{q(x)}{p(x)} \right] \nu(dx) \tag{17.27}$$

- In summation form, we have

$$D(q||p) = \sum_{x \in \mathcal{X}^m} q(x) \left[\log \frac{q(x)}{p(x)} \right] \tag{17.28}$$

- For exponential models this takes on some interesting forms, and more over, we can see the variational approximation above as a KL-divergence minimization problem.

Mean field, KL-Divergence, Exponential Model Families

- Given two distributions p, q, KL-Divergence of p w.r.t. q is defined as

$$D(q||p) = \int_{\mathcal{X}^m} q(x) \left[\log \frac{q(x)}{p(x)} \right] \nu(dx) \tag{17.27}$$

- In summation form, we have

$$D(q||p) = \sum_{x \in \mathcal{X}^m} q(x) \left[\log \frac{q(x)}{p(x)} \right] \tag{17.28}$$

- For exponential models this takes on some interesting forms, and more over, we can see the variational approximation above as a KL-divergence minimization problem.
- Recall, exponential models can be parameterized using canonical parameters θ or mean parameters μ. We will use notational shortcuts: $D(\theta^1||\theta^2) \equiv D(p_{\theta^1}||p_{\theta^2})$, and $D(\mu^1||\mu^2) \equiv D(p_{\mu^1}||p_{\mu^2})$, and even $D(\mu^1||\theta^2) \equiv D(p_{\mu^1}||p_{\theta^2})$.

Mean field, KL-Divergence, Exponential Model Families

- Consider $\theta^1, \theta^2 \in \Omega$

Mean field, KL-Divergence, Exponential Model Families

- Consider $\theta^1, \theta^2 \in \Omega$
- Let $D(\theta^1 || \theta^2)$ have aforementioned meaning (KL-divergence between the two corresponding distributions), and let $\mu^i = \mathbb{E}_{\theta^i}[\phi(X)]$,

Mean field, KL-Divergence, Exponential Model Families

- Consider $\theta^1, \theta^2 \in \Omega$
- Let $D(\theta^1 || \theta^2)$ have aforementioned meaning (KL-divergence between the two corresponding distributions), and let $\mu^i = \mathbb{E}_{\theta^i}[\phi(X)]$,
- Then we have a Bregman divergence form:

$$D(\theta^1 || \theta^2) = \mathbb{E}_{\theta^1}\left[\log \frac{p_{\theta^1}(x)}{p_{\theta^2}(x)}\right] \tag{17.29}$$

$$= A(\theta^2) - A(\theta^1) - \left\langle \mu^1, \theta^2 - \theta^1 \right\rangle \tag{17.30}$$

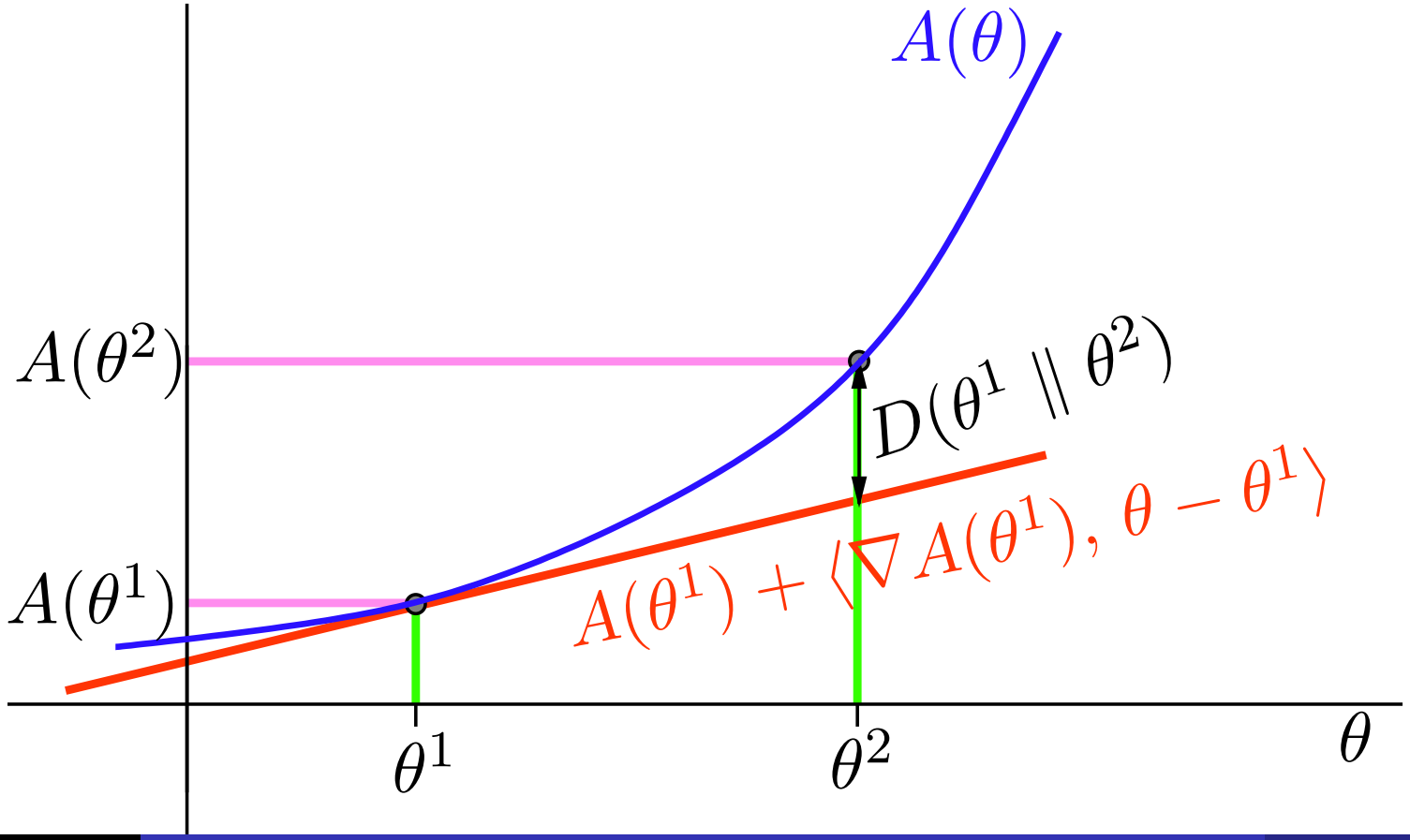
$$= A(\theta^2) - \left[A(\theta^1) + \left\langle \nabla A(\theta^1), \theta^2 - \theta^1 \right\rangle\right] \tag{17.31}$$

where $\mu^1 = \nabla A(\theta^1)$ can be seen as the gradient/slope of $A(\theta)$ evaluated at θ^1.

Mean field, KL-Divergence, Exponential Model Families

$$D(\theta^1||\theta^2) = A(\theta^2) - A(\theta^1) - \langle \mu^1, \theta^2 - \theta^1 \rangle \tag{17.32}$$

$$= A(\theta^2) - \left[A(\theta^1) + \langle \nabla A(\theta^1), \theta^2 - \theta^1 \rangle \right] \tag{17.33}$$



Mean field, KL-Divergence, Exponential Model Families

- We can also express a mixed/hybrid form of KL in terms of dual $A^*(\mu) = \sup_{\theta \in \Omega} (\langle \theta, \mu \rangle - A(\theta)) \geq \langle \theta', \mu \rangle - A(\theta')$ for any $\theta' \in \Omega$.

Mean field, KL-Divergence, Exponential Model Families

- We can also express a mixed/hybrid form of KL in terms of dual $A^*(\mu) = \sup_{\theta \in \Omega} (\langle \theta, \mu \rangle - A(\theta)) \geq \langle \theta', \mu \rangle - A(\theta')$ for any $\theta' \in \Omega$.
- We can also write the KL as:

$$D(\theta^1 || \theta^2) = A(\theta^2) - A(\theta^1) - \left\langle \mu^1, \theta^2 - \theta^1 \right\rangle \tag{17.34}$$

$$= A(\theta^2) - \left\langle \mu^1, \theta^2 \right\rangle - \left[A(\theta^1) - \left\langle \mu^1, \theta^1 \right\rangle \right] \tag{17.35}$$

$$= A(\theta^2) - \left\langle \mu^1, \theta^2 \right\rangle + A^*(\mu^1) \triangleq D(\mu^1 || \theta^2) \tag{17.36}$$

which comes from dual expression $A^*(\mu^1) = \left\langle \theta^1, \mu^1 \right\rangle - A(\theta^1)$ which holds for the dually coupled parameters $\mu^1 = \mathbb{E}_{\theta^1}[\phi(X)]$.

Mean field, KL-Divergence, Exponential Model Families

- We can also express a mixed/hybrid form of KL in terms of dual $A^*(\mu) = \sup_{\theta \in \Omega} (\langle \theta, \mu \rangle - A(\theta)) \geq \langle \theta', \mu \rangle - A(\theta')$ for any $\theta' \in \Omega$.
- We can also write the KL as:

$$D(\theta^1 || \theta^2) = A(\theta^2) - A(\theta^1) - \left\langle \mu^1, \theta^2 - \theta^1 \right\rangle \tag{17.34}$$

$$= A(\theta^2) - \left\langle \mu^1, \theta^2 \right\rangle - \left[A(\theta^1) - \left\langle \mu^1, \theta^1 \right\rangle \right] \tag{17.35}$$

$$= A(\theta^2) - \left\langle \mu^1, \theta^2 \right\rangle + A^*(\mu^1) \triangleq D(\mu^1 || \theta^2) \tag{17.36}$$

  which comes from dual expression $A^*(\mu^1) = \left\langle \theta^1, \mu^1 \right\rangle - A(\theta^1)$ which holds for the dually coupled parameters $\mu^1 = \mathbb{E}_{\theta^1}[\phi(X)]$.
- In particular, this equation (variational expression for the cumulant):

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{ \langle \theta, \mu \rangle - A^*(\mu) \} \tag{13.7}$$

Mean field, KL-Divergence, Exponential Model Families

- We can also express a mixed/hybrid form of KL in terms of dual $A^*(\mu) = \sup_{\theta\in\Omega}(\langle\theta,\mu\rangle - A(\theta)) \geq \langle\theta',\mu\rangle - A(\theta')$ for any $\theta' \in \Omega$.
- We can also write the KL as:

$$D(\theta^1||\theta^2) = A(\theta^2) - A(\theta^1) - \left\langle\mu^1, \theta^2 - \theta^1\right\rangle \tag{17.34}$$

$$= A(\theta^2) - \left\langle\mu^1, \theta^2\right\rangle - \left[A(\theta^1) - \left\langle\mu^1, \theta^1\right\rangle\right] \tag{17.35}$$

$$= A(\theta^2) - \left\langle\mu^1, \theta^2\right\rangle + A^*(\mu^1) \triangleq D(\mu^1||\theta^2) \tag{17.36}$$

  which comes from dual expression $A^*(\mu^1) = \left\langle\theta^1, \mu^1\right\rangle - A(\theta^1)$ which holds for the dually coupled parameters $\mu^1 = \mathbb{E}_{\theta^1}[\phi(X)]$.
- In particular, this equation (variational expression for the cumulant):

$$A(\theta) = \sup_{\mu\in\mathcal{M}} \{\langle\theta,\mu\rangle - A^*(\mu)\} \tag{13.7}$$

- ...can be written as:

$$\inf_{\mu\in\mathcal{M}} \{A(\theta) + A^*(\mu) - \langle\theta,\mu\rangle\} = \inf_{\mu\in\mathcal{M}} D(\mu||\theta) = 0 \tag{17.37}$$

Mean field, KL-Divergence, Exponential Model Families

- Thus, solving the mean-field variational problem (see Eqn. (17.26)) of:

$$\max_{\mu \in \mathcal{M}_F(G)} \{\langle \mu, \theta \rangle - A_F^*(\mu)\} = \max_{\mu \in \mathcal{M}_F(G)} \{\langle \mu, \theta \rangle - A^*(\mu)\} \quad (17.38)$$

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

Mean field, KL-Divergence, Exponential Model Families

- Thus, solving the mean-field variational problem (see Eqn. (17.26)) of:

$$\max_{\mu \in \mathcal{M}_F(G)} \{\langle \mu, \theta \rangle - A_F^*(\mu)\} = \max_{\mu \in \mathcal{M}_F(G)} \{\langle \mu, \theta \rangle - A^*(\mu)\} \quad (17.38)$$

  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

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

Naïve Mean field for Ising Model

- A classic example of mean-field (goes back to statistical physics)
- Mean parameters for Ising: $\mu_s = \mathbb{E}[X_s] = p(X_s = 1)$, $\mu_{st} = \mathbb{E}[X_s X_t] = p(X_s = 1, X_t = 1)$, thus $\mu \in \mathbb{R}^{|V|+|E|}$.

Naïve Mean field for Ising Model

- A classic example of mean-field (goes back to statistical physics)
- Mean parameters for Ising: $\mu_s = \mathbb{E}[X_s] = p(X_s = 1)$, $\mu_{st} = \mathbb{E}[X_s X_t] = p(X_s = 1, X_t = 1)$, thus $\mu \in \mathbb{R}^{|V|+|E|}$.
- Let $F_0 = (V, \emptyset)$ be our mean field approximation family. Thus,

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

Naïve Mean field for Ising Model

- A classic example of mean-field (goes back to statistical physics)
- Mean parameters for Ising: $\mu_s = \mathbb{E}[X_s] = p(X_s = 1)$, $\mu_{st} = \mathbb{E}[X_s X_t] = p(X_s = 1, X_t = 1)$, thus $\mu \in \mathbb{R}^{|V|+|E|}$.
- Let $F_0 = (V, \emptyset)$ be our mean field approximation family. Thus,

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

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

$$-A^*_{F_0}(\mu) = \sum_{s \in V} H_s(\mu_s) \tag{17.39}$$

which are sum of unary entropy terms, very cheap.

Naïve Mean field for Ising Model

- A classic example of mean-field (goes back to statistical physics)
- Mean parameters for Ising: $\mu_s = \mathbb{E}[X_s] = p(X_s = 1)$, $\mu_{st} = \mathbb{E}[X_s X_t] = p(X_s = 1, X_t = 1)$, thus $\mu \in \mathbb{R}^{|V|+|E|}$.
- Let $F_0 = (V, \emptyset)$ be our mean field approximation family. Thus,

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

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

$$-A^*_{F_0}(\mu) = \sum_{s \in V} H_s(\mu_s) \tag{17.39}$$

  which are sum of unary entropy terms, very cheap.
- Moreover, polytope for $M_{F_0}(G)$ is also very simple, namely the hypercube $[0, 1]^m$.

Naive Mean field for Ising Model

- We get variational lower bound problem

$$A(\theta) \geq \max_{(\mu_1,\ldots,\mu_m)\in[0,1]^m} \left\{ \sum_{s\in V} \theta_s \mu_s + \sum_{(s,t)\in E} \theta_{st}\mu_s\mu_t + \sum_{s\in V} H_s(\mu_s) \right\} \tag{17.40}$$

Naive Mean field for Ising Model

- We get variational lower bound problem

$$A(\theta) \geq \max_{(\mu_1,\ldots,\mu_m)\in[0,1]^m} \left\{ \sum_{s\in V} \theta_s \mu_s + \sum_{(s,t)\in E} \theta_{st}\mu_s\mu_t + \sum_{s\in V} H_s(\mu_s) \right\} \tag{17.40}$$

- Have constrained form of edge mean parameters $\mu_{st} = \mu_s \mu_t$

Naive Mean field for Ising Model

- We get variational lower bound problem

$$A(\theta) \geq \max_{(\mu_1,\ldots,\mu_m)\in[0,1]^m} \left\{ \sum_{s\in V} \theta_s \mu_s + \sum_{(s,t)\in E} \theta_{st}\mu_s\mu_t + \sum_{s\in V} H_s(\mu_s) \right\} \tag{17.40}$$

- 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.

Naive Mean field for Ising Model

- We get variational lower bound problem

$$A(\theta) \geq \max_{(\mu_1,\ldots,\mu_m)\in[0,1]^m} \left\{ \sum_{s\in V} \theta_s \mu_s + \sum_{(s,t)\in E} \theta_{st}\mu_s\mu_t + \sum_{s\in V} H_s(\mu_s) \right\} \tag{17.40}$$

- 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.
- Once again, we have a non-convex problem.

Naive Mean field for Ising Model

- We get variational lower bound problem

$$A(\theta) \geq \max_{(\mu_1,\ldots,\mu_m)\in[0,1]^m} \left\{ \sum_{s\in V} \theta_s \mu_s + \sum_{(s,t)\in E} \theta_{st}\mu_s\mu_t + \sum_{s\in V} H_s(\mu_s) \right\} \tag{17.40}$$

- 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.
- Once again, we have a non-convex problem.
- One way to optimize is to do coordinate ascent (given otherwise fixed vector, optimize one value at a time).

Naive Mean field for Ising Model

- We get variational lower bound problem

$$A(\theta) \geq \max_{(\mu_1,\ldots,\mu_m)\in[0,1]^m} \left\{ \sum_{s\in V} \theta_s \mu_s + \sum_{(s,t)\in E} \theta_{st}\mu_s\mu_t + \sum_{s\in V} H_s(\mu_s) \right\} \tag{17.40}$$

- 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.
- Once again, we have a non-convex problem.
- One way to optimize is to do coordinate ascent (given otherwise fixed vector, optimize one value at a time).
- If each coordinate optimization is optimal, we'll get a stationary point.

Naive Mean field for Ising Model

- We get variational lower bound problem

$$A(\theta) \geq \max_{(\mu_1,\ldots,\mu_m)\in[0,1]^m} \left\{ \sum_{s\in V} \theta_s \mu_s + \sum_{(s,t)\in E} \theta_{st}\mu_s\mu_t + \sum_{s\in V} H_s(\mu_s) \right\} \tag{17.40}$$

- 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.
- Once again, we have a non-convex problem.
- One way to optimize is to do coordinate ascent (given otherwise fixed vector, optimize one value at a time).
- If each coordinate optimization is optimal, we'll get a stationary point.
- Fortunately, each coordinate optimization is concave!

Naive Mean field for Ising Model

- coordinate ascent: choose some s and optimize μ_s fixing all μ_t for $t \neq s$.

Naive Mean field for Ising Model

- coordinate ascent: choose some s and optimize μ_s fixing all μ_t for $t \neq s$.
- Taking derivatives w.r.t. μ_s, we get the following update rule for element μ_s

$$\mu_s \leftarrow \sigma\left(\theta_s + \sum_{t \in N(s)} \theta_{st}\mu_t\right) \tag{17.41}$$

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

Naive Mean field for Ising Model

- coordinate ascent: choose some s and optimize μ_s fixing all μ_t for $t \neq s$.
- Taking derivatives w.r.t. μ_s, we get the following update rule for element μ_s

$$\mu_s \leftarrow \sigma\left(\theta_s + \sum_{t \in N(s)} \theta_{st}\mu_t\right) \tag{17.41}$$

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

- This is the classic mean-field update that is quite well known, but derived from coordinate assent optimization of a variational perspective of the problem.

Naive Mean field for Ising Model

- coordinate ascent: choose some s and optimize μ_s fixing all μ_t for $t \neq s$.
- Taking derivatives w.r.t. μ_s, we get the following update rule for element μ_s

$$\mu_s \leftarrow \sigma\left(\theta_s + \sum_{t \in N(s)} \theta_{st}\mu_t\right) \tag{17.41}$$

  where $\sigma(z) = [1 + \exp(-z)]^{-1}$ is the sigmoid (logistic) function.
- This is the classic mean-field update that is quite well known, but derived from coordinate assent optimization of a variational perspective of the problem.
- The variational approach indeed seems quite general and powerful.

Example of Lack of Convexity

- Consider simple two variable example (X_1, X_2), $X_i \in \{-1, +1\}$.

Example of Lack of Convexity

- Consider simple two variable example (X_1, X_2), $X_i \in \{-1, +1\}$.
- Exponential family form

$$p_\theta(x) \propto \exp(\theta_1 x_1 + \theta_2 x_2 + \theta_{12} x_1 x_2) \tag{17.42}$$

having mean parameters $\mu_i = \mathbb{E}[X_i]$ and $\mu_{12} = \mathbb{E}[X_1 X_2]$.

Example of Lack of Convexity

- Consider simple two variable example (X_1, X_2), $X_i \in \{-1, +1\}$.
- Exponential family form

$$p_\theta(x) \propto \exp(\theta_1 x_1 + \theta_2 x_2 + \theta_{12} x_1 x_2) \tag{17.42}$$

  having mean parameters $\mu_i = \mathbb{E}[X_i]$ and $\mu_{12} = \mathbb{E}[X_1 X_2]$.
- Impose constraint $\mu_{12} = \mu_1 \mu_2$, we get mean field objective

$$f(\mu_1, \mu_2; \theta) = \theta_{12}\mu_1\mu_2 + \theta_1\mu_1 + \theta_2\mu_2 + H(\mu_1) + H(\mu_2) \tag{17.43}$$

  where $H(\mu_i) = -\frac{1}{2}(1+\mu_i)\log\frac{1}{2}(1+\mu_i) - \frac{1}{2}(1-\mu_i)\log\frac{1}{2}(1-\mu_i)$
  Note that $p(X_i = +1) = \frac{1}{2}(1+\mu_i)$

Example of Lack of Convexity

- Consider simple two variable example (X_1, X_2), $X_i \in \{-1, +1\}$.
- Exponential family form

$$p_\theta(x) \propto \exp(\theta_1 x_1 + \theta_2 x_2 + \theta_{12} x_1 x_2) \tag{17.42}$$

  having mean parameters $\mu_i = \mathbb{E}[X_i]$ and $\mu_{12} = \mathbb{E}[X_1 X_2]$.
- Impose constraint $\mu_{12} = \mu_1 \mu_2$, we get mean field objective

$$f(\mu_1, \mu_2; \theta) = \theta_{12}\mu_1\mu_2 + \theta_1\mu_1 + \theta_2\mu_2 + H(\mu_1) + H(\mu_2) \tag{17.43}$$

  where $H(\mu_i) = -\frac{1}{2}(1+\mu_i)\log\frac{1}{2}(1+\mu_i) - \frac{1}{2}(1-\mu_i)\log\frac{1}{2}(1-\mu_i)$
- Consider sub-models of the form:

$$(\theta_1, \theta_2, \theta_{12}) = \left(0, 0, \frac{1}{4}\log\frac{q}{1-q}\right) \triangleq \theta(q) \tag{17.44}$$

  where $q \in (0,1)$ is a parameter such that, for any q we have $\mathbb{E}[X_i] = 0$. It turns out that in this form, we have $q = p(X_1 = X_2)$.

Example of Lack of Convexity

- Consider simple two variable example (X_1, X_2), $X_i \in \{-1, +1\}$.
- Exponential family form

$$p_\theta(x) \propto \exp(\theta_1 x_1 + \theta_2 x_2 + \theta_{12} x_1 x_2) \tag{17.42}$$

  having mean parameters $\mu_i = \mathbb{E}[X_i]$ and $\mu_{12} = \mathbb{E}[X_1 X_2]$.
- Impose constraint $\mu_{12} = \mu_1 \mu_2$, we get mean field objective

$$f(\mu_1, \mu_2; \theta) = \theta_{12}\mu_1\mu_2 + \theta_1\mu_1 + \theta_2\mu_2 + H(\mu_1) + H(\mu_2) \tag{17.43}$$

  where $H(\mu_i) = -\frac{1}{2}(1+\mu_i)\log\frac{1}{2}(1+\mu_i) - \frac{1}{2}(1-\mu_i)\log\frac{1}{2}(1-\mu_i)$
- Consider sub-models of the form:

$$(\theta_1, \theta_2, \theta_{12}) = \left(0, 0, \frac{1}{4}\log\frac{q}{1-q}\right) \triangleq \theta(q) \tag{17.44}$$

  where $q \in (0,1)$ is a parameter such that, for any q we have $\mathbb{E}[X_i] = 0$. It turns out that in this form, we have $q = p(X_1 = X_2)$.
- Is mean field objective in this case convex for all q?

Lack of Convexity example

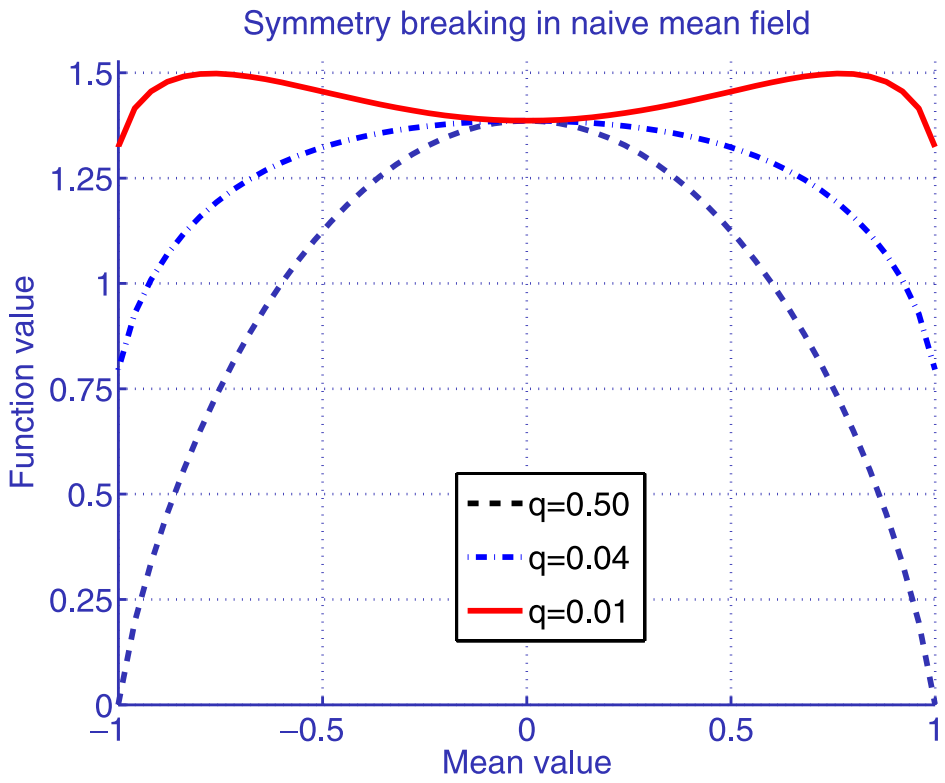
- For $q = 0.5$, objective $f(\mu_1, \mu_2; \theta(0.5))$ has global maximum at $(\mu_1, \mu_2) = (0, 0)$ so mean field is exact and convex. This corresponds to $p(X_1 = X_2) = 0$.

Lack of Convexity example

- For $q = 0.5$, objective $f(\mu_1, \mu_2; \theta(0.5))$ has global maximum at $(\mu_1, \mu_2) = (0, 0)$ so mean field is exact and convex. This corresponds to $p(X_1 = X_2) = 0$.
- When q gets small, f becomes non-convex, e.g., has multiple modes in figure.

Lack of Convexity example

- For $q = 0.5$, objective $f(\mu_1, \mu_2; \theta(0.5))$ has global maximum at $(\mu_1, \mu_2) = (0, 0)$ so mean field is exact and convex. This corresponds to $p(X_1 = X_2) = 0$.
- When q gets small, f becomes non-convex, e.g., has multiple modes in figure.



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

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 $\mathcal{M}(F)$ be 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 $\mathcal{M}(F)$ be 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 $\mathcal{M}(F)$ be 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 $\mathcal{M}(F)$ be 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 $\mathcal{M}(F)$ be 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.
- Example: mean field Ising, $\mu_{st} = g(\mu(F)) = \mu_s \mu_t$.

Structured Mean Field

- The mean field optimization problem becomes

$$\max_{\mu \in \mathcal{M}_F(G)} \{\langle \mu, \theta \rangle - A_F^*(\mu)\} \tag{17.45}$$

$$= \max_{\mu(F) \in \mathcal{M}(F)} \underbrace{\left\{ \sum_{\alpha \in \mathcal{I}(F)} \theta_\alpha \mu_\alpha + \sum_{\alpha \in \mathcal{I}^c(F)} \theta_\alpha g_\alpha(\mu(F)) - A_F^*(\mu(F)) \right\}}_{f(\mu(F))} \tag{17.46}$$

Structured Mean Field

- The mean field optimization problem becomes

$$\max_{\mu \in \mathcal{M}_F(G)} \{\langle \mu, \theta \rangle - A_F^*(\mu)\} \tag{17.45}$$

$$= \max_{\mu(F) \in \mathcal{M}(F)} \underbrace{\left\{ \sum_{\alpha \in \mathcal{I}(F)} \theta_\alpha \mu_\alpha + \sum_{\alpha \in \mathcal{I}^c(F)} \theta_\alpha g_\alpha(\mu(F)) - A_F^*(\mu(F)) \right\}}_{f(\mu(F))} \tag{17.46}$$

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

$$\frac{\partial f}{\partial \mu_\beta}(\mu(F)) = \theta_\beta + \sum_{\alpha \in \mathcal{I}(G) \setminus \mathcal{I}(F)} \theta_\alpha \frac{\partial g_\alpha}{\partial \mu_\beta}(\mu(F)) - \frac{\partial A_F^*}{\partial \mu_\beta}(\mu(F)) \tag{17.47}$$

Structured Mean Field

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

$$\nabla A_F^*(\mu(F)) = \theta + \sum_{\alpha \in \mathcal{I}(G) \setminus \mathcal{I}(F)} \theta_\alpha \nabla g_\alpha(\mu(F)) \tag{17.48}$$

Structured Mean Field

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

$$\nabla A_F^*(\mu(F)) = \theta + \sum_{\alpha \in \mathcal{I}(G) \setminus \mathcal{I}(F)} \theta_\alpha \nabla g_\alpha(\mu(F)) \tag{17.48}$$

- ∇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 $\beta \in \mathcal{I}(F)$

$$\gamma_\beta(F) \leftarrow \theta_\beta + \sum_{\alpha \in \mathcal{I}(G) \setminus \mathcal{I}(F)} \theta_\alpha \frac{\partial g_\alpha}{\partial \mu_\beta}(\mu(F)) \tag{17.49}$$

Structured Mean Field

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

$$\nabla A_F^*(\mu(F)) = \theta + \sum_{\alpha \in \mathcal{I}(G) \setminus \mathcal{I}(F)} \theta_\alpha \nabla g_\alpha(\mu(F)) \tag{17.48}$$

- ∇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 $\beta \in \mathcal{I}(F)$

$$\gamma_\beta(F) \leftarrow \theta_\beta + \sum_{\alpha \in \mathcal{I}(G) \setminus \mathcal{I}(F)} \theta_\alpha \frac{\partial g_\alpha}{\partial \mu_\beta}(\mu(F)) \tag{17.49}$$

- Above is the mean field update, mapping from a canonical parameters (θ_β for $\beta \in \mathcal{I}(F)$) 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. (17.49), 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. (17.49), 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. (17.49).

Structured Mean Field

- Alternatively, we can transform back to mean parameters right away using ∇A is the forward mapping, maping from mean to canonical.

Structured Mean Field

- Alternatively, we can transform back to mean parameters right away using ∇A is the forward mapping, maping from mean to canonical.
- 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 mean to canonical.
- 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 $\beta \in \mathcal{I}(F)$:

$$\mu_\beta(F) \leftarrow \frac{\partial A_F}{\partial \gamma_\beta}\left(\theta_\beta + \sum_{\alpha \in \mathcal{I}(G)\setminus\mathcal{I}(F)} \theta_\alpha \nabla g_\alpha(\mu(F))\right) \tag{17.50}$$

Structured Mean Field

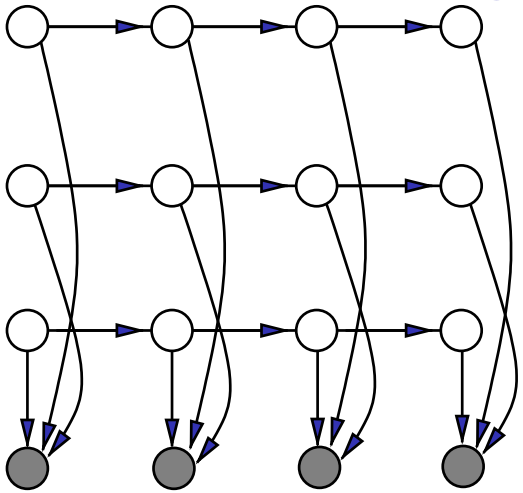
- Alternatively, we can transform back to mean parameters right away using ∇A is the forward mapping, maping from mean to canonical.
- I.e., we can derive a mean field mean parameter to mean parameter update equation using A_F since $\nabla A_F(\gamma(F)) = \mu(F)$,
- We get update, for $\beta \in \mathcal{I}(F)$:

$$\mu_\beta(F) \leftarrow \frac{\partial A_F}{\partial \gamma_\beta}\left(\theta_\beta + \sum_{\alpha \in \mathcal{I}(G)\setminus\mathcal{I}(F)} \theta_\alpha \nabla g_\alpha(\mu(F))\right) \tag{17.50}$$

- This generalizes our mean field coordinate ascent update from before, where in that case we would have $\frac{\partial A_F}{\partial \gamma_\beta}$ being the sigmoid mapping.

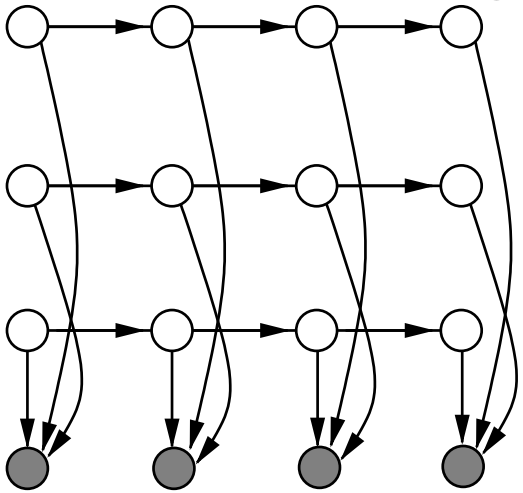
Structured Mean Field Factorial HMMs

- This idea was developed and applied using factorial HMMs.



Structured Mean Field Factorial HMMs

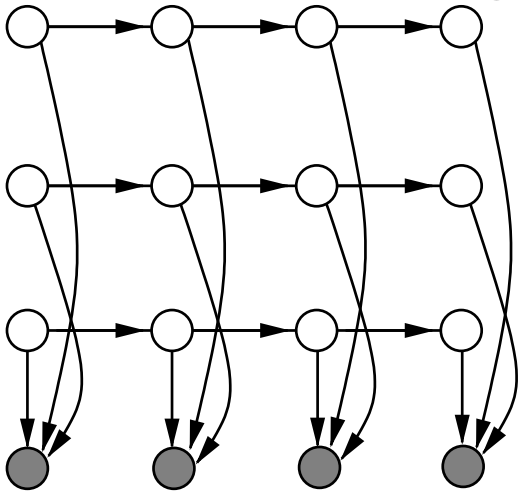
- This idea was developed and applied using factorial HMMs.



- Graph consists of M 1st-order Markov chains $y^i_{1:T}$ 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

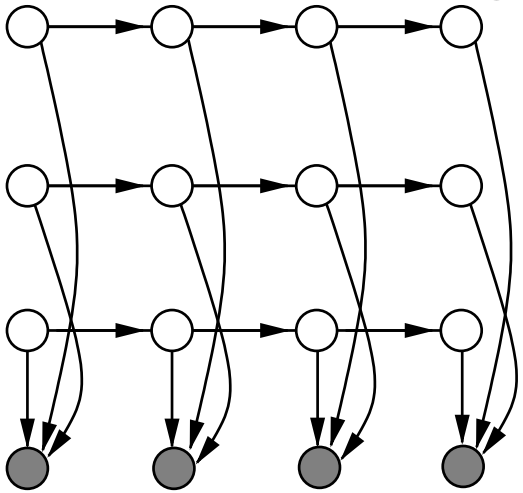
- This idea was developed and applied using factorial HMMs.



- Graph consists of M 1st-order Markov chains $y^i_{1:T}$ 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, if there are M chains, we have a clique of size M.

Structured Mean Field Factorial HMMs

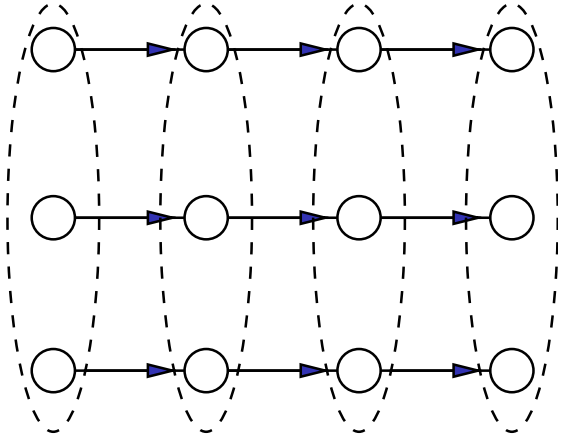
- This idea was developed and applied using factorial HMMs.



- Graph consists of M 1st-order Markov chains $y^i_{1:T}$ 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, if there are M chains, we have a clique of size M.
- Here, 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 $E \setminus F = \left\{ \left\{ x^1_t, x^2_t, \ldots, x^M_t \right\} : t \in [T] \right\}$.

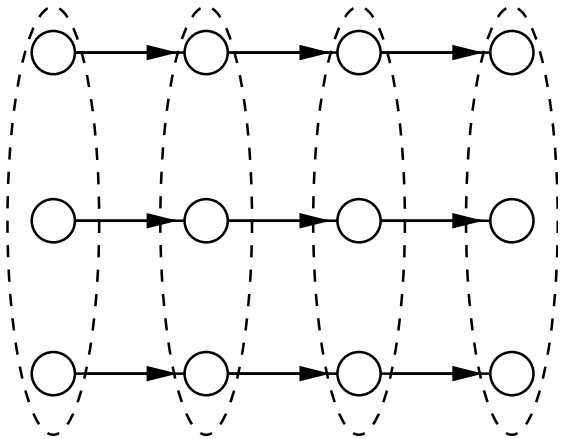
Structured Mean Field Factorial HMMs

- The induced dependencies (cliques as dotted ellipses)



Structured Mean Field Factorial HMMs

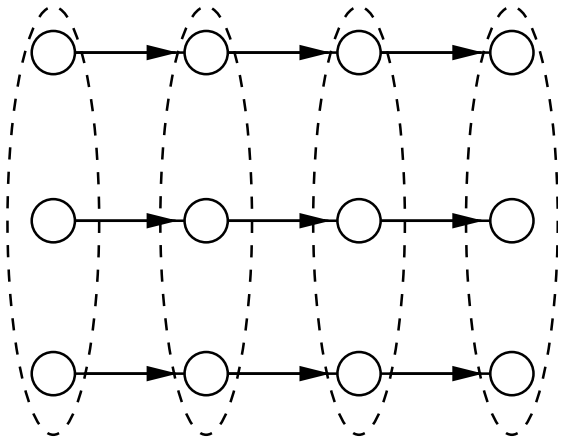
- The induced dependencies (cliques as dotted ellipses)



- Tree width of this model is?

Structured Mean Field Factorial HMMs

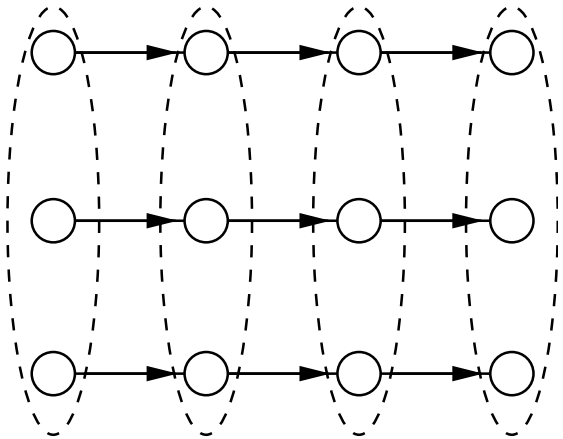
- The induced dependencies (cliques as dotted ellipses)



- Tree width of this model is? M

Structured Mean Field Factorial HMMs

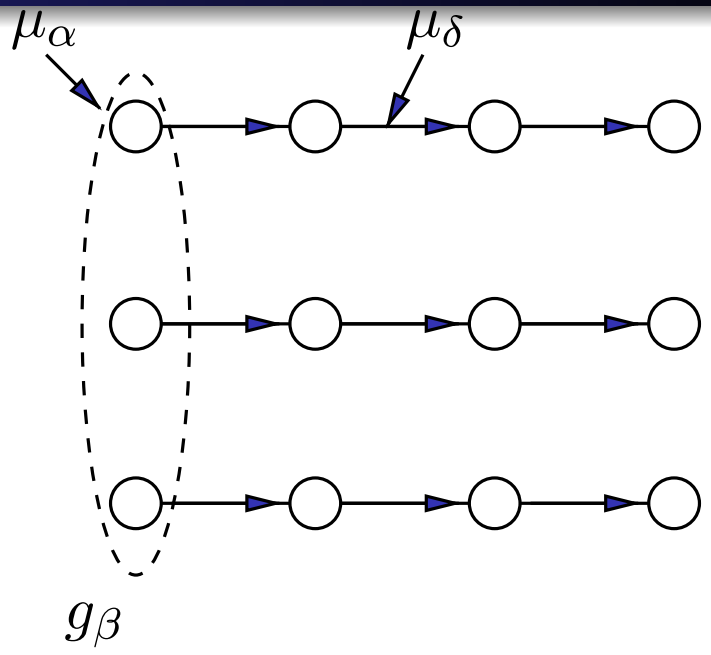
- The induced dependencies (cliques as dotted ellipses)



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

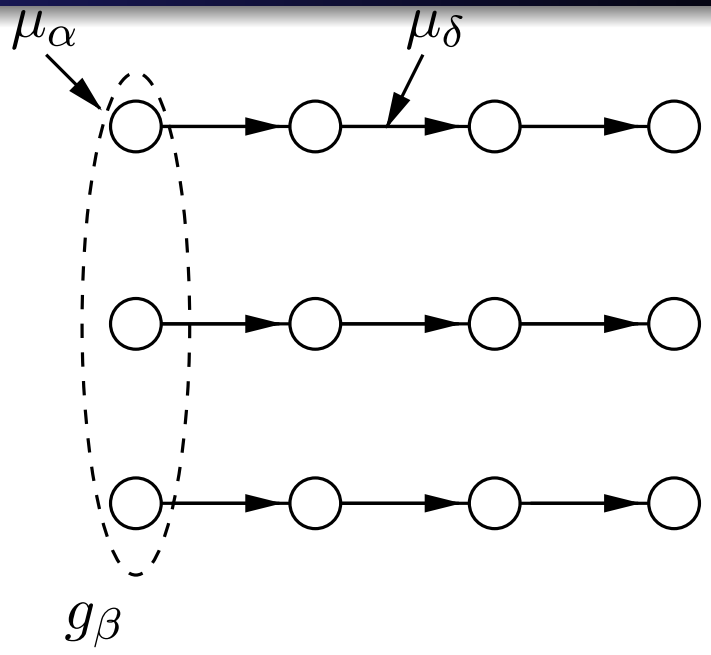
Structured Mean Field Factorial HMMs

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



Structured Mean Field Factorial HMMs

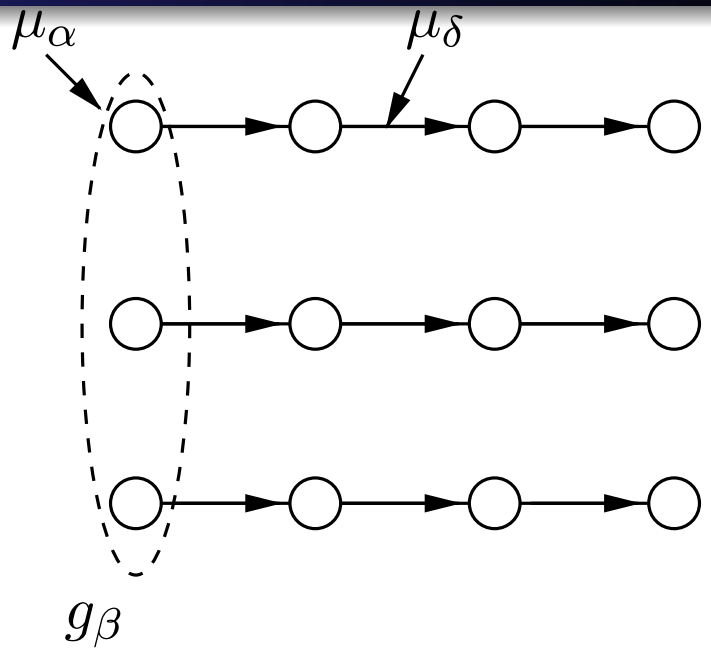
- A "natural" choice of approximating distribution is a set of coupled chains, natural, perhaps primarily for computational reasons.



- 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

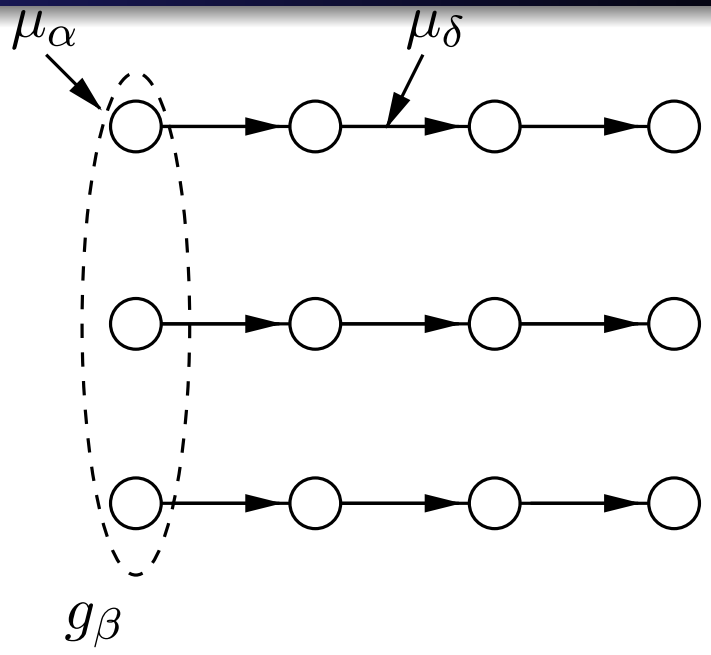
- A "natural" choice of approximating distribution is a set of coupled chains, natural, perhaps primarily for computational reasons.



- 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. (17.49) 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. (17.49) updates parameters for $\beta \in \mathcal{I}(F)$, corresponds to all edges of all M Markov chains.
- To recover mean parameters (or do Eqn. (17.50)), need only forward-backward procedure on each chain separately, $O(MTr^2)$.

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

$$A(\theta) \geq \max_{\mu \in \mathcal{M}_F(G)} \{\langle \mu, \theta \rangle - A_F^*(\mu)\} = A_{\text{mf}}(\theta) \tag{17.1}$$

Convex Relaxations and Upper Bounds

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{17.51}$$

- What about upper bounds?

Convex Relaxations and Upper Bounds

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{17.51}$$

- What about upper bounds?
- Other than mean field, none of the other approximation methods have been anything other than approximation methods.

Convex Relaxations and Upper Bounds

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{17.51}$$

- What about upper bounds?
- Other than mean field, none of the other approximation methods have been anything other than approximation methods.
- We would like both lower and upper bounds of $A(\theta)$ since that will allow us to produce upper and lower bounds of the probabilistic queries we wish to perform.

Convex Relaxations and Upper Bounds

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{17.51}$$

- What about upper bounds?
- Other than mean field, none of the other approximation methods have been anything other than approximation methods.
- We would like both lower and upper bounds of $A(\theta)$ since that will allow us to produce upper and lower bounds of the probabilistic queries we wish to perform.
- If the upper and lower bounds between a given probably p is small, $p_L \leq p \leq p_U$, with $p_U - p_L \leq \epsilon$, we have guarantees, for a particular instance of a model.

Convex Relaxations and Upper Bounds

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{17.51}$$

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

Convex Relaxations and Upper Bounds - Relaxed Entropy

- 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) 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) 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) 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) 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) 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 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) 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 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, so that $\mu(F) \in \mathcal{M}(F)$. Thus, $\mathcal{M}(F) \subseteq \mathbb{R}^{|\mathcal{I}(F)|}$, and

$$\mathcal{M}(F) = \left\{ \mu \in \mathbb{R}^{|\mathcal{I}(F)|} | \exists p \text{ s.t. } \mu_\alpha = \mathbb{E}_p[\phi_\alpha(X)] \; \forall \alpha \in \mathcal{I}(F) \right\} \tag{17.52}$$

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 17.6.1

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

$$A^*(\mu(F)) \leq A^*(\mu) \tag{17.53}$$

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

- Intuition: $H(\mu) = H(p_\mu)$ is the entropy of the exponential family model with mean parameters μ.

Convex Relaxations and Upper Bounds - Relaxed Entropy

Proposition 17.6.1

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

$$A^*(\mu(F)) \leq A^*(\mu) \tag{17.53}$$

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

- Intuition: $H(\mu) = H(p_\mu)$ is the entropy of the exponential family model with mean parameters 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 17.6.1

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

$$A^*(\mu(F)) \leq A^*(\mu) \tag{17.53}$$

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

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

Convex Relaxations and Upper Bounds - Relaxed Entropy

Proposition 17.6.1

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

$$A^*(\mu(F)) \leq A^*(\mu) \tag{17.53}$$

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

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

Convex Relaxations and Upper Bounds - Relaxed Entropy

Proof.

- Dual problem

$$A^*(\mu) = \sup_{\theta \in \mathbb{R}^d} \{\langle \mu, \theta \rangle - A(\theta)\} \tag{17.54}$$

Convex Relaxations and Upper Bounds - Relaxed Entropy

Proof.

- Dual problem

$$A^*(\mu) = \sup_{\theta \in \mathbb{R}^d} \{\langle \mu, \theta \rangle - A(\theta)\} \tag{17.54}$$

- Dual problem in sub-graph case.

$$A^*(\mu(F)) = \sup_{\theta(F) \in \mathbb{R}^{d(F)}} \{\langle \mu(F), \theta(F) \rangle - A(\theta(F))\} \tag{17.55}$$

Convex Relaxations and Upper Bounds - Relaxed Entropy

Proof.

- Dual problem

$$A^*(\mu) = \sup_{\theta \in \mathbb{R}^d} \{\langle \mu, \theta \rangle - A(\theta)\} \tag{17.54}$$

- Dual problem in sub-graph case.

$$A^*(\mu(F)) = \sup_{\theta(F) \in \mathbb{R}^{d(F)}} \{\langle \mu(F), \theta(F) \rangle - A(\theta(F))\} \tag{17.55}$$

- Dual problem in sub-graph case — alternate expression

$$A^*(\mu(F)) = \sup_{\substack{\theta \in \mathbb{R}^d \\ \theta_\alpha = 0 \; \forall \alpha \notin \mathcal{I}(F)}} \{\langle \mu, \theta \rangle - A(\theta)\} \tag{17.56}$$

Convex Relaxations and Upper Bounds - Relaxed Entropy

Proof.

- Dual problem

$$A^*(\mu) = \sup_{\theta \in \mathbb{R}^d} \{\langle \mu, \theta \rangle - A(\theta)\} \tag{17.54}$$

- Dual problem in sub-graph case.

$$A^*(\mu(F)) = \sup_{\theta(F) \in \mathbb{R}^{d(F)}} \{\langle \mu(F), \theta(F) \rangle - A(\theta(F))\} \tag{17.55}$$

- Dual problem in sub-graph case — alternate expression

$$A^*(\mu(F)) = \sup_{\substack{\theta \in \mathbb{R}^d \\ \theta_\alpha = 0 \; \forall \alpha \notin \mathcal{I}(F)}} \{\langle \mu, \theta \rangle - A(\theta)\} \tag{17.56}$$

- 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, gives general upper bound

$$H(\mu) \leq \mathbb{E}_\rho[H(\mu(F))] = \sum_{F \in \mathfrak{D}} \rho(F) H(\mu(F)) \tag{17.57}$$

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, gives general upper bound

$$H(\mu) \leq \mathbb{E}_\rho[H(\mu(F))] = \sum_{F \in \mathfrak{D}} \rho(F) H(\mu(F)) \tag{17.57}$$

- This will be our convexified upper bound on entropy.

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, gives general upper bound

$$H(\mu) \leq \mathbb{E}_\rho[H(\mu(F))] = \sum_{F \in \mathfrak{D}} \rho(F) H(\mu(F)) \tag{17.57}$$

- This will be our convexified upper bound on entropy.
- compared to mean field, we are not choosing only one structure, but many of them, and mixing them together in certain ways.

Convex Relaxations and Upper Bounds - Outer bound

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

Convex Relaxations and Upper Bounds - Outer bound

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

$$\mathcal{L}(G;\mathfrak{D}) = \left\{\tau \in \mathbb{R}^d | \tau(F) \in \mathcal{M}(F) \;\; \forall F \in \mathfrak{D}\right\} \tag{17.58}$$

$$= \bigcap_{F \in \mathfrak{D}} \mathcal{M}(F) \tag{17.59}$$

Convex Relaxations and Upper Bounds - Outer bound

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

$$\mathcal{L}(G;\mathfrak{D}) = \left\{\tau \in \mathbb{R}^d | \tau(F) \in \mathcal{M}(F) \;\; \forall F \in \mathfrak{D}\right\} \tag{17.58}$$

$$= \bigcap_{F \in \mathfrak{D}} \mathcal{M}(F) \tag{17.59}$$

- 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 $\mathcal{M}(F)$ (i.e., non-neg, sums to 1, and consistency).

Convex Relaxations and Upper Bounds - Outer bound

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

$$\mathcal{L}(G;\mathfrak{D}) = \left\{\tau \in \mathbb{R}^d | \tau(F) \in \mathcal{M}(F) \;\; \forall F \in \mathfrak{D}\right\} \tag{17.58}$$

$$= \bigcap_{F \in \mathfrak{D}} \mathcal{M}(F) \tag{17.59}$$

- 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 $\mathcal{M}(F)$ (i.e., non-neg, sums to 1, and consistency).
- 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.

$$B_{\mathfrak{D}}(\theta;\rho) \triangleq \sup_{\tau \in \mathcal{L}(G;\mathfrak{D})} \left\{ \langle \tau, \theta \rangle + \sum_{F \in \mathfrak{D}} \rho(F) H(\tau(F)) \right\} \tag{17.60}$$

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.

$$B_{\mathfrak{D}}(\theta;\rho) \triangleq \sup_{\tau\in\mathcal{L}(G;\mathfrak{D})} \left\{ \langle \tau, \theta \rangle + \sum_{F\in\mathfrak{D}} \rho(F) H(\tau(F)) \right\} \tag{17.60}$$

- 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.

$$B_{\mathfrak{D}}(\theta;\rho) \triangleq \sup_{\tau\in\mathcal{L}(G;\mathfrak{D})} \left\{ \langle \tau, \theta \rangle + \sum_{F\in\mathfrak{D}} \rho(F) H(\tau(F)) \right\} \tag{17.60}$$

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

$$B_{\mathfrak{D}}(\theta;\rho) \triangleq \sup_{\tau\in\mathcal{L}(G;\mathfrak{D})} \left\{ \langle \tau, \theta \rangle + \sum_{F\in\mathfrak{D}} \rho(F) H(\tau(F)) \right\} \tag{17.60}$$

- Objective is convex in θ since it is a max over a set of affine functions of θ (i.e., $g(\theta) = \max_\tau \langle \tau, \theta \rangle + c_\tau$)
- 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, $A(\theta) \leq B_{\mathfrak{D}}(\theta;\rho)$

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`