## 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 he intracxtable ones (of which there are $d_{I}$ ). Thus, $d=d_{T}+d_{I}$.
- Tractable component

$$
\begin{equation*}
\phi \triangleq\left(\phi_{1}, \phi_{2}, \ldots, \phi_{d_{T}}\right) \tag{17.5}
\end{equation*}
$$

- Intractable component

$$
\begin{equation*}
\Phi \triangleq\left(\Phi^{1}, \Phi^{2}, \ldots, \Phi^{d_{I}}\right) \tag{17.6}
\end{equation*}
$$

- $\phi_{i}$ are typically univariate, while $\Phi^{i}$ 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 $(\phi, \Phi)$.


## Associated Distributions: base and $i$-augmented

- The associated exponential family

$$
\begin{align*}
p(x ; \theta, \tilde{\theta}) & \propto \exp (\langle\theta, \phi(x)\rangle) \exp (\langle\tilde{\theta}, \Phi(x)\rangle)  \tag{17.7}\\
& =\exp (\langle\theta, \phi(x)\rangle) \prod_{i=1}^{d_{I}} \exp \left(\left\langle\tilde{\theta}^{i}, \Phi^{i}(x)\right\rangle\right) \tag{17.8}
\end{align*}
$$

- Base model is tractable

$$
\begin{equation*}
p(x ; \theta \widehat{0}) \gamma \exp (\langle\theta, \phi(x)\rangle) \tag{17.9}
\end{equation*}
$$

- $\Phi^{i}$-augmented model

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

## New EP-based outer bound

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

$$
\begin{equation*}
\Pi^{i}(\tau, \tilde{\tau}) \rightarrow\left(\tau, \tilde{\tau}^{i}\right) \tag{17.14}
\end{equation*}
$$

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)

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

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

$$
\begin{align*}
\tau \in \mathcal{M}(\phi) & \Leftrightarrow \exists p \text { s.t. } \tau=E_{p}[\phi(X)]  \tag{17.16}\\
(\tau, \tilde{\tau}) \in \mathcal{L}(\phi, \Phi) & \Leftrightarrow \tau \in \mathcal{M}(\phi) \& \& \mid \exists \text { s.t. }\left(\tau, \tilde{\tau}^{i}\right) \neq E_{p}\left[\phi(X), \Phi^{i}(X)\right] \\
(\tau, \tilde{\tau}) \in \mathcal{M}(\phi, \Phi) & \Leftrightarrow \exists p \text { s.t. }(\tau, \tilde{\tau})=E_{p}[\phi(X), \Phi(X)]
\end{align*}
$$

- If $\Phi^{i}$ 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 $\phi$-exponential family which mean parameters $\tau$ with entropy $H(\tau)$; B$)$ Also, for $i=1 \ldots d_{I}$, there is a member of the $\left(\phi, \Phi^{i}\right)$-exponential family with mean parameters $\left(\tau, \tilde{\tau}^{i}\right)$ with entropy $H\left(\tau, \tilde{\tau}^{i}\right)$.
- Both entropy forms are easy to compute, and so is a new entropy approximation:

$$
\begin{equation*}
H(\tau, \tilde{\tau}) \approx H_{\mathrm{ep}}(\tau, \tilde{\tau}) \triangleq H(\tau)+\sum_{\ell=1}^{d_{I}}\left[H\left(\tau, \tilde{\tau}^{l}\right)-H(\tau)\right] \tag{17.14}
\end{equation*}
$$

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

$$
\begin{equation*}
\max _{(\tau, \tilde{\tau}) \in \mathcal{L}(\phi, \Phi)}\left\{\langle\tau, \theta\rangle+\langle\tilde{\tau}, \tilde{\theta}\rangle+H_{\mathrm{ep}}(\tau, \tilde{\tau})\right\} \tag{17.15}
\end{equation*}
$$

- This characterizes the EP algorithms.
- Given graph $G=(V, E)$ when we take $\phi$ to be unaries $V$ and $\Phi$ 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 $i \in\left[d_{T}\right]$.
- This gives large set of pseudo-mean parameters

$$
\begin{equation*}
\left\{\tau,\left(\eta^{i}, \tilde{\tau}^{i}\right), i \in\left[d_{I}\right]\right\} \in \mathbb{R}^{d_{T}} \times\left(\mathbb{R}^{d_{T}} \times \mathbb{R}^{b}\right)^{d_{I}} \tag{17.14}
\end{equation*}
$$

- We arrive at the optimization:
$\max _{\left\{\tau,\left\{\left(\eta^{i}, \tilde{\tau}^{i}\right)\right\}_{i}\right\}}\left\{\langle\tau, \theta\rangle+\sum_{i=1}^{d_{I}}\left\langle\tilde{\tau}^{i}, \tilde{\theta}^{i}\right\rangle+H(\tau)+\sum_{i=1}^{d_{I}}\left[H\left(\eta^{i}, \tilde{\tau}^{i}\right)-H\left(\eta^{i}\right)\right]\right\}$
(17.15)
subject to $\tau \in \mathcal{M}(\phi)$, and for all $i$ that $\tau=\eta^{i}$ and that $\left(\eta^{i}, \tilde{\tau}^{i}\right) \in \mathcal{M}\left(\phi, \Phi^{i}\right)$.
- Use Lagrange multipliers to impose constraint $\eta^{i}=\tau$ for all $i$, and for the rest of the constraints too.


## Moment Matching $\rightarrow$ Expectation Propagation Updates

(1) At iteration $n=0$, initialize the Lagrange multiplier vectors $\left(\lambda^{1}, \ldots, \lambda^{d_{I}}\right)$
(2) At each iteration $n=1,2, \ldots$ choose someqinglex $i(n) \in\left\{1, \ldots, \alpha_{\not}\right\}$.
(3) Under the following augnented distribution

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

compute the mean parameters $\eta^{i}$ as follows:

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

(9) Form base distribution $q$ using Equation ?? and adjust $\lambda^{i(2)}$ to satisfy the moment-matching condition

$$
\begin{equation*}
\mathbb{E}_{q}[\phi(X)]=\eta^{i(n)} \tag{17.21}
\end{equation*}
$$

(6) 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

$$
\begin{equation*}
A(\theta)=\sup _{\mu \in \mathcal{M}}\left\{\langle\theta, \mu\rangle-A^{*}(\mu)\right\} \tag{17.1}
\end{equation*}
$$

where dual takes form:

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

## Variational Approach Amenable to Approximation

- Original variational representation of $\log$ partition function

$$
\begin{equation*}
A(\theta)=\sup _{\mu \in \mathcal{M}}\left\{\langle\theta, \mu\rangle-A^{*}(\mu)\right\} \tag{17.1}
\end{equation*}
$$

where dual takes form:

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

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


## Variational Approach Amenable to Approximation

- Original variational representation of $\log$ partition function

$$
\begin{equation*}
A(\theta)=\sup _{\mu \in \mathcal{M}}\left\{\langle\theta, \mu\rangle-A^{*}(\mu)\right\} \tag{17.1}
\end{equation*}
$$

where dual takes form:

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

- 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 $H_{\text {app }}=\sum_{g \in E} c(g) H_{g}\left(\tau_{g}\right)$ (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 $H_{\text {app }}=\sum_{g \in E} c(g) H_{g}\left(\tau_{g}\right)$ (via Möbius) to get Kikuchi variational approximation, message passing on hypergraphs.
(3) Partition $\tau$ into $(\tau, \tilde{\tau})$, and set $\mathcal{M} \leftarrow \mathcal{L}(\phi, \Phi)$ and set $-A^{*}(\mu) \leftarrow H_{\text {ep }}(\tau, \tilde{\tau})$ to get expectation propagation.

## 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 $\phi_{s}\left(x_{s}\right)$ for $s \in V$ and each edge $\phi_{u, v}\left(x_{u}, x_{v}\right)$ 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 $\phi_{s}\left(x_{s}\right)$ for $s \in V$ and each edge $\phi_{u, v}\left(x_{u}, x_{v}\right)$ for $(u, v) \in E(G)$.
- Base distribution is only the nodes (fully factored independent distribuiton)

$$
\begin{equation*}
p\left(x ; \phi_{1}, \ldots, \phi_{m}, \overrightarrow{0}\right) \propto \prod_{v \in V} \exp \left(\theta_{s}\left(x_{s}\right)\right) \tag{17.1}
\end{equation*}
$$

## 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 $\phi_{s}\left(x_{s}\right)$ for $s \in V$ and each edge $\phi_{u, v}\left(x_{u}, x_{v}\right)$ for $(u, v) \in E(G)$.
- Base distribution is only the nodes (fully factored independent distribuiton)

$$
\begin{equation*}
p\left(x ; \phi_{1}, \ldots, \phi_{m}, \overrightarrow{0}\right) \propto \prod_{v \in V} \exp \left(\theta_{s}\left(x_{s}\right)\right) \tag{17.1}
\end{equation*}
$$

- Each $\Phi^{i}$ 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:

$$
\begin{equation*}
p\left(x ; \phi_{1}, \ldots, \phi_{m}, \phi_{u v}\right) \propto \prod_{v \in V} \exp \left(\theta_{s}\left(x_{s}\right)\right) \exp \left(\theta_{u v}\left(x_{u}, x_{v}\right)\right) \tag{17.2}
\end{equation*}
$$

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

- Base entropy is sum of node marginal entropies

$$
\begin{equation*}
H\left(T_{1}, \ldots, \tau_{m}\right)=\sum_{\ell \in V} H\left(\sigma_{s}\right) \tag{17.3}
\end{equation*}
$$

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

- Base entropy is sum of node marginal entropies

$$
\begin{equation*}
H\left(\tau_{1}, \ldots, \tau_{m}\right)=\sum_{s \in V} H\left(\tau_{s}\right) \tag{17.3}
\end{equation*}
$$

- Augmented entropy takes the form

where $I\left(\tau_{u, v}\right)$ is the mutual information between $X_{u}$ and $X_{v}$ under joint distribution $\tau_{u v}$.


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

- Base entropy is sum of node marginal entropies

$$
\begin{equation*}
H\left(\tau_{1}, \ldots, \tau_{m}\right)=\sum_{s \in V} H\left(\tau_{s}\right) \tag{17.3}
\end{equation*}
$$

- Augmented entropy takes the form

$$
\begin{align*}
H\left(\tau_{1}, \ldots, \tau_{m}, \tau_{u v}\right) & =\sum_{s \in V \backslash\{u, v\}} H\left(\tau_{s}\right)+H\left(\tau_{u v}\right)  \tag{17.4}\\
& =\sum_{s \in V} H\left(\tau_{s}\right)+\left[H\left(\tau_{u v}\right)-H\left(\tau_{u}\right)-H\left(\tau_{v}\right)\right] \\
& =\sum_{s \in V} H\left(\tau_{s}\right)+I\left(\tau_{u, v}\right) \tag{17.6}
\end{align*}
$$

where $I\left(\tau_{u, v}\right)$ is the mutual information between $X_{u}$ and $X_{v}$ under joint distribution $\tau_{u v}$.

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

$$
\begin{equation*}
H_{\mathrm{ep}}(\tau)=\sum_{s \in V} H\left(\tau_{s}\right) \not \sum_{(u, v) \in E(G)} I\left(\tau_{u v}\right) \tag{17.7}
\end{equation*}
$$

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

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


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

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


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

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


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

- the base mean parameter $\mathcal{M}(\phi)$ just asks that $\tau=\left(\tau_{s}, s \in V\right)$ are valid unary marginals (i.e., non-negative and sum to one, in the form of $\forall s \in V, 0 \leq \tau_{s}\left(x_{s}\right) \leq 1$ and $\sum_{x_{s}} \tau_{s}\left(x_{s}\right)=1$.
- Each augmentation $\mathcal{M}\left(\phi, \Phi^{u v}\right)$ for edge $(u, v) \in E(G)$ also asks that $\tau_{u v}$ marginalizes down to $\tau_{u}$ and $\tau_{v}$, i.e., $\sum_{x_{v}} \tau_{u v}\left(x_{v}, x_{u}\right)=\tau_{u}\left(x_{u}\right)$ and $\sum_{x_{u}} \tau_{u v}\left(x_{v}, x_{u}\right)=\tau_{v}\left(x_{v}\right)$.
- Then considering $\mathcal{L}(\phi, \Phi)$ as defined, we must have for all $(u, v) \in E(G), \Pi^{u v}(\tau, \tilde{\tau}) \in \mathcal{M}\left(\phi, \Phi^{u v}\right)$ - 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:

$$
\begin{align*}
q(x ; \theta, \lambda) & \propto \prod_{s \in V} \exp \left(\theta_{s}\left(x_{s}\right)\right) \prod_{(u, v) \in E} \exp \left(\lambda_{u v}\left(x_{v}\right)+\lambda_{v u}\left(x_{u}\right)\right) \quad \text { (17.8) } \\
& =\prod_{s \in V} \exp \left(\theta_{s}\left(x_{s}\right)+\sum_{t \in N(s)} \frac{\left.\lambda_{t s}\left(x_{s}\right)\right)}{v_{1}}\right.  \tag{17.9}\\
& \propto \prod_{s \in V} \tau_{s}\left(x_{s}\right) \\
\text { where } \tau_{s}\left(x_{s}\right) & =\exp \left(\theta_{s}\left(x_{s}\right)+\sum_{t \in N(s)} \lambda_{t s}\left(x_{s}\right)\right)
\end{align*}
$$

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

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

$$
\left.\begin{array}{l}
\qquad \begin{array}{rl}
q(x ; \theta, \lambda) & \propto \prod_{s \in V} \exp \left(\theta_{s}\left(x_{s}\right)\right) \prod_{(u, v) \in E} \exp \left(\lambda_{u v}\left(x_{v}\right)+\lambda_{v u}\left(x_{u}\right)\right) \\
& =\prod_{s \in V} \exp \left(\theta_{s}\left(x_{s}\right)+\sum_{t \in N(s)} \lambda_{t s}\left(x_{s}\right)\right) \\
& \propto \prod_{s \in V} \tau_{s}\left(x_{s}\right)
\end{array} \\
\text { where } \tau_{s}\left(x_{s}\right)
\end{array}\right)=\exp \left(\theta_{s}\left(x_{s}\right)+\sum_{t \in N(s)} \lambda_{t s}\left(x_{s}\right)\right) . ~ \$
$$

- 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 $\ell=(u, v)$,

$$
q^{(u, v)}(x ; \theta, \lambda) \propto q(x ; \theta, \lambda) \exp \left(\theta_{u v}\left(x_{u}, x_{v}\right)-\lambda_{u v}\left(x_{v}\right)-\lambda_{u v}\left(x_{u}\right)\right)
$$

$$
\left.=\prod \prod_{-v} \tau_{s}\left(x_{s}\right)\right] \exp \left(\theta_{u v}\left(x_{u}, x_{v}\right)-\lambda_{u v}\left(x_{v}\right)-\lambda_{u v}\left(x_{u}\right)\right)
$$

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

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

$$
\begin{align*}
q^{(u, v)}(x ; \theta, \lambda) & \propto q(x ; \theta, \lambda) \exp \left(\theta_{u v}\left(x_{u}, x_{v}\right)-\lambda_{u v}\left(x_{v}\right)-\lambda_{u v}\left(x_{u}\right)\right) \\
& =\left[\prod_{s \in V} \tau_{s}\left(x_{s}\right)\right] \exp \left(\theta_{u v}\left(x_{u}, x_{v}\right)-\lambda_{u v}\left(x_{v}\right)-\lambda_{u v}\left(x_{u}\right)\right) \tag{17.11}
\end{align*}
$$

- 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 $\lambda_{u v}\left(x_{v}\right)$ and $\lambda_{v u}\left(x_{u}\right)$ in Equation (17.8) so that the marginal distributions $\tau_{v}\left(x_{v}\right)$ and $\tau_{u}\left(x_{u}\right)$ match the marginals of the joint along this edge.


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

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

$$
\begin{align*}
q^{(u, v)}(x ; \theta, \lambda) & \propto q(x ; \theta, \lambda) \exp \left(\theta_{u v}\left(x_{u}, x_{v}\right)-\lambda_{u v}\left(x_{v}\right)-\lambda_{\Psi v}\left(x_{u}\right)\right) \\
& =\left[\prod_{s \in V} \tau_{s}\left(x_{s}\right)\right] \exp \left(\theta_{u v}\left(x_{u}, x_{v}\right)-\lambda_{u v}\left(x_{v}\right)-\lambda_{\psi v}\left(x_{u}\right)\right) \tag{17.11}
\end{align*}
$$

- 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 $\lambda_{u v}\left(x_{v}\right)$ and $\lambda_{v u}\left(x_{u}\right)$ in Equation (17.8) so that the marginal distributions $\tau_{v}\left(x_{v}\right)$ and $\tau_{u}\left(x_{u}\right)$ 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 augm 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:

$$
\begin{equation*}
p(x ; \theta, \overrightarrow{0}) \propto \prod_{s \in V} \exp \left(\theta_{s}\left(x_{s}\right)\right) \prod_{(s, t) \in E(T)} \exp \left(\theta_{s t}\left(x_{s}, x_{t}\right)\right) \tag{17.12}
\end{equation*}
$$

## 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:

$$
\begin{equation*}
p(x ; \theta, \overrightarrow{0}) \propto \prod_{s \in V} \exp \left(\theta_{s}\left(x_{s}\right)\right) \prod_{(s, t) \in E(T)} \exp \left(\theta_{s t}\left(x_{s}, x_{t}\right)\right) \tag{17.12}
\end{equation*}
$$

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

$$
\begin{equation*}
p\left(x ; \theta, \theta_{u, v}\right) \operatorname{\propto p}(x ; \theta, \overrightarrow{0}) \exp \left(\theta_{u, v}\left(x_{u}, x_{v}\right)\right) \tag{17.13}
\end{equation*}
$$

## 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 $\Phi^{i}$ 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 $\Phi^{i}$ 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 $\Phi^{i}$ 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 $\Phi^{i}$ 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 $\Phi^{i}$ 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 $H_{\text {app }}=\sum_{g \in E} c(g) H_{g}\left(\tau_{g}\right)$ (via Möbius) to get Kikuchi variational approximation, message passing on hypergraphs.
(3) Partition $\tau$ into $(\tau, \tilde{\tau})$, and set $\mathcal{M} \leftarrow \mathcal{L}(\phi, \Phi)$ and set $-A^{*}(\mu) \leftarrow H_{\text {ep }}(\tau, \tilde{\tau})$ to get expectation propagation.

## Mean Field

- 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, $F=\left(V, E_{F}\right)$ 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, $F=\left(V, E_{F}\right)$ 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, $F=\left(V, E_{F}\right)$ 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, $F=\left(V, E_{F}\right)$ where $E_{F} \subseteq E$.
- Simplest example: $F=(V, \emptyset)$ all independence model.
- Tree example: $F=\left(V, E_{T}\right)$ 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, $F=\left(V, E_{F}\right)$ where $E_{F} \subseteq E$.
- Simplest example: $F=(V, \emptyset)$ all independence model.
- Tree example: $F=\left(V, E_{T}\right)$ where edges $E_{T} \subset E$ constitute a spanning tree.
- Exponential family, sufficient statistics $\phi=\left(\phi_{\alpha}, \alpha \in \mathcal{I}\right)$ 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, $F=\left(V, E_{F}\right)$ where $E_{F} \subseteq E$.
- Simplest example: $F=(V, \emptyset)$ all independence model.
- Tree example: $F=\left(V, E_{T}\right)$ where edges $E_{T} \subset E$ constitute a spanning tree.
- Exponential family, sufficient statistics $\phi=\left(\phi_{\alpha}, \alpha \in \mathcal{I}\right)$ associated with this family $\mathcal{I}(F) \subseteq \mathcal{I}$. These are the statistics that need respect the Markov properties of only the subgraph $F$.
- $\Omega$ gets smaller too. The parameters that respect $F$ are of the form:

$$
\begin{equation*}
\mathbb{R}^{|\mathcal{I}|} \ni \Omega(F) \triangleq\left\{\theta \in \Omega \mid \theta_{\alpha}=0 \quad \forall \alpha \in \mathcal{I} \backslash \mathcal{I}(F)\right\} \subseteq \Omega \tag{17.14}
\end{equation*}
$$

## 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, $F=\left(V, E_{F}\right)$ where $E_{F} \subseteq E$.
- Simplest example: $F=(V, \emptyset)$ all independence model.
- Tree example: $F=\left(V, E_{T}\right)$ where edges $E_{T} \subset E$ constitute a spanning tree.
- Exponential family, sufficient statistics $\phi=\left(\phi_{\alpha}, \alpha \in \mathcal{I}\right)$ associated with this family $\mathcal{I}(F) \subseteq \mathcal{I}$. These are the statistics that need respect the Markov properties of only the subgraph $F$.
- $\Omega$ gets smaller too. The parameters that respect $F$ are of the form:

$$
\begin{equation*}
\mathbb{R}^{|\mathcal{I}|} \ni \Omega(F) \triangleq\left\{\theta \in \Omega \mid \theta_{\alpha}=0 \quad \forall \alpha \in \mathcal{I} \backslash \mathcal{I}(F)\right\} \subseteq \Omega \tag{17.14}
\end{equation*}
$$

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, $F=\left(V, E_{F}\right)$ where $E_{F} \subseteq E$.
- Simplest example: $F=(V, \emptyset)$ all independence model.
- Tree example: $F=\left(V, E_{T}\right)$ where edges $E_{T} \subset E$ constitute a spanning tree.
- Exponential family, sufficient statistics $\phi=\left(\phi_{\alpha}, \alpha \in \mathcal{I}\right)$ associated with this family $\mathcal{I}(F) \subseteq \mathcal{I}$. These are the statistics that need respect the Markov properties of only the subgraph $F$.
- $\Omega$ gets smaller too. The parameters that respect $F$ are of the form:

$$
\begin{equation*}
\mathbb{R}^{|\mathcal{I}|} \ni \Omega(F) \triangleq\left\{\theta \in \Omega \mid \theta_{\alpha}=0 \quad \forall \alpha \in \mathcal{I} \backslash \mathcal{I}(F)\right\} \subseteq \Omega \tag{17.14}
\end{equation*}
$$

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

$$
\begin{equation*}
\Omega\left(F_{0}\right)=\left\{\theta \in \Omega \mid \theta_{(s, t)}=0 \quad \forall(s, t) \in E(G)\right\} \tag{17.15}
\end{equation*}
$$

## 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

$$
\begin{equation*}
\Omega\left(F_{0}\right)=\left\{\theta \in \Omega \mid \theta_{(s, t)}=0 \quad \forall(s, t) \in E(G)\right\} \tag{17.15}
\end{equation*}
$$

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

$$
\begin{equation*}
p_{\theta}(x)=\prod_{s \in V} p\left(x_{s} ; \theta_{s}\right) \tag{17.16}
\end{equation*}
$$

## 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

$$
\begin{equation*}
\Omega\left(F_{0}\right)=\left\{\theta \in \Omega \mid \theta_{(s, t)}=0 \quad \forall(s, t) \notin T\right\} \tag{17.17}
\end{equation*}
$$

## 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

$$
\begin{equation*}
\Omega\left(F_{0}\right)=\left\{\theta \in \Omega \mid \theta_{(s, t)}=0 \quad \forall(s, t) \notin T\right\} \tag{17.17}
\end{equation*}
$$

- This gives a tree-dependent family

$$
\begin{equation*}
p_{\theta}(x)=\prod_{s \in V} p\left(x_{s} ; \theta_{s}\right) \prod_{(s, t) \in T} \frac{p\left(x_{s}, x_{t} ; \theta_{s t}\right)}{p\left(x_{s} ; \theta_{s}\right) p\left(x_{t} ; \theta_{t}\right)} \tag{17.18}
\end{equation*}
$$

## Inner bound Approximate Polytope

- Before, we had $\mathcal{M}(G ; \phi)\left(=\mathcal{M}_{G}(G ; \phi)\right)$, all possible mean parameters associated with $G$ and associafed set of sufficient statistics $\phi$.


## Inner bound Approximate Polytope

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

$$
\begin{align*}
& \mathcal{M}_{F}(G ; \phi)=\left\{\mu \in \mathbb{R}^{d} \mid \mu=\mathbb{E}_{\theta}[\phi(x)] \text { for some } \theta \in \Omega(F)\right\}  \tag{17.19}\\
& \bigcap \bigcap \bigcap
\end{align*}
$$

## Inner bound Approximate Polytope

- Before, we had $\mathcal{M}(G ; \phi)\left(=\mathcal{M}_{G}(G ; \phi)\right)$, all possible mean parameters associated with $G$ and associated set of sufficient statistics $\phi$.
- For a given subgraph $F$, we only consider those mean parameters possible under $F$-respecting models. I.e.,
$\mathcal{M}_{F}(G ; \phi)=\left\{\mu \in \mathbb{R}^{d} \mid \mu=\mathbb{E}_{\theta}[\phi(x)]\right.$ for som $\theta \in \Omega(F)$
- Therefore, since $\theta \in \Omega(F) \subseteq \Omega$, we have that

and so $\mathcal{M}_{F}^{\circ}(G ; \phi)$ is ann er approximation of the set of realizable mean parameters.


## Inner bound Approximate Polytope

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

$$
\begin{equation*}
\mathcal{M}_{F}(G ; \phi)=\left\{\mu \in \mathbb{R}^{d} \mid \mu=\mathbb{E}_{\theta}[\phi(x)] \text { for some } \theta \in \Omega(F)\right\} \tag{17.19}
\end{equation*}
$$

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

$$
\mathcal{M}_{F}^{\circ}(G ; \phi) \subseteq \mathcal{M}^{\circ}(G ; \phi)
$$


(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:

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

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

## Mean field variational lower bound

## Proof.

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


## Mean field variational lower bound

## Proof.

- On the one hand, obvious due to $A(\theta)=\sup _{\mu \in \mathcal{M}}\left\{\langle\theta, \mu\rangle-A^{*}(\mu)\right\}$
- More tradit proof, let be any distribution that satisfies moment match $\operatorname{EE}_{q}[\phi(X)]=\mu$, then:
$A(\theta)=\log \int_{\mathcal{X}^{m}} \exp \langle\theta, \phi(x)\rangle \nu(d x)$
(17.22)
(17.23)
(17.24)
(17.25)


## Mean field variational lower bound

## Proof.

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

$$
\begin{align*}
A(\theta) & =\log \int_{\mathcal{X}^{m}} \exp \langle\theta, \phi(x)\rangle \nu(d x)  \tag{17.22}\\
& =\log \int_{\mathcal{X}^{m}} q(x) \frac{\exp \langle\theta, \phi(x)\rangle}{q(x)} \nu(d x)  \tag{17.23}\\
& \geq \int_{\mathcal{X}^{m}} q(x)[\langle\theta, \phi(x)\rangle-\log q(x)] \nu(d x)  \tag{17.24}\\
& =\left\langle\theta, E_{q}[\phi(X)]\right\rangle-H(q)=\langle\theta, \mu\rangle-H(q) \tag{17.25}
\end{align*}
$$

- 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 $A(\theta)=\sup _{\mu \in \mathcal{M}}\left\{\langle\theta, \mu\rangle-A^{*}(\mu)\right\}$
- More traditional proof, let $q$ be any distribution that satisfies moment matching $\mathbb{E}_{q}[\phi(X)]=\mu$, then:

$$
\begin{align*}
A(\theta) & =\log \int_{\mathcal{X}^{m}} \exp \langle\theta, \phi(x)\rangle \nu(d x)  \tag{17.22}\\
& =\log \int_{\mathcal{X}^{m}} q(x) \frac{\exp \langle\theta, \phi(x)\rangle}{q(x)} \nu(d x)  \tag{17.23}\\
& \geq \int_{\mathcal{X}^{m}} q(x)[\langle\theta, \phi(x)\rangle-\log q(x)] \nu(d x)  \tag{17.24}\\
& =\left\langle\theta, E_{q}[\phi(X)]\right\rangle-H(q)=\langle\theta, \mu\rangle-H(q) \tag{17.25}
\end{align*}
$$

- 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 $A_{\mathrm{MF}}(\theta)$ where:

$$
\begin{equation*}
A(\theta) \geq \max _{\mu \in \mathcal{M}_{F}(G)}\left\{\langle\mu, \theta\rangle-A_{F}^{*}(\mu)\right\} \triangleq A_{\mathrm{MF}}(\theta) \tag{17.26}
\end{equation*}
$$

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

## 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 $A_{\mathrm{MF}}(\theta)$ where:

$$
\begin{equation*}
A(\theta) \geq \max _{\mu \in \mathcal{M}_{F}(G)}\left\{\langle\mu, \theta\rangle-A_{F}^{*}(\mu)\right\} \triangleq A_{\mathrm{MF}}(\theta) \tag{17.26}
\end{equation*}
$$

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

- Note, for $\mu \in \mathcal{M}_{F}(G)$ 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:
(b) Partition function has variational representation (dual of dual)

$$
\begin{equation*}
\widehat{A(\theta)}=\sup _{\mu \in \mathcal{M}}\left\{\langle\theta, \mu\rangle-A^{*}(\mu)\right\} \tag{17.4}
\end{equation*}
$$

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

## 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

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

## Mean field, KL-Divergence, Exponential Model Families

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

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

- In summation form, we have

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

## Mean field, KL-Divergence, Exponential Model Families

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

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

- In summation form, we have

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

- 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

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

- In summation form, we have

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

- 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 $\theta$ or mean parameters $\mu$. We will use notational shortcuts:
$D\left(\theta^{1} \| \theta^{2}\right) \equiv D\left(p_{\theta^{1}} \| p_{\theta^{2}}\right)$, and $D\left(\mu^{1} \| \mu^{2}\right)=D\left(p_{\mu^{1}} \| p_{\mu^{2}}\right)$, and even $D\left(\mu^{1}| | \theta^{2}\right) \equiv D\left(p_{\mu^{1}} \mid p_{\theta^{2}}\right)$.


## 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 $D\left(\theta^{1} \| \theta^{2}\right)$ 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 $D\left(\theta^{1} \| \theta^{2}\right)$ 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:

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

$$
\begin{aligned}
& h_{\theta^{\prime}}\left(\theta^{\prime}\right)=A\left(\theta^{\prime}\right) \\
& h_{\theta^{\prime}}(\theta) \geq A(\theta)
\end{aligned}
$$

## Mean field, KL-Divergence, Exponential Model Families

$$
\begin{align*}
D\left(\theta^{1} \| \theta^{2}\right) & =A\left(\theta^{2}\right)-A\left(\theta^{1}\right)-\left\langle\mu^{1}, \theta^{2}-\theta^{1}\right\rangle  \tag{17.32}\\
& =A\left(\theta^{2}\right)-\left[A\left(\theta^{1}\right)+\left\langle\nabla A\left(\theta^{1}\right), \theta^{2}-\theta^{1}\right\rangle\right] \tag{17.33}
\end{align*}
$$



## Mean field, KL-Divergence, Exponential Model Families

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


## Mean field, KL-Divergence, Exponential Model Families

- We can also express a mixed/hybrid form of KL terms of dual $A^{*}(\mu)=\sup _{\theta \in \Omega}(\langle\theta, \mu\rangle-A(\theta)) \not\left\langle\theta^{\prime}, \mu\right\rangle-A\left(\theta^{\prime}\right)$ for any $\theta^{\prime} \in \Omega$.
- We can also write the KL as:

$$
\begin{align*}
D\left(\theta^{1} \| \theta^{2}\right) & =A\left(\theta^{2}\right)-A\left(\theta^{1}\right)-\left\langle\mu^{1}, \theta^{2}-\theta^{1}\right\rangle  \tag{17.34}\\
& =A\left(\theta^{2}\right)-\left\langle\mu^{1}, \theta^{2}\right)-\left[A\left(\theta^{1}\right)-\left\langle\mu^{1}, \theta^{1}\right\rangle\right] \\
& =A\left(\theta^{2}\right)-\left\langle\mu^{1}, \theta^{2}\right\rangle+A^{*}\left(\mu^{1}\right) \triangleq D\left(\mu^{1} \| \theta^{2}\right)
\end{align*}
$$

which comes from dual expression $A^{*}\left(\mu^{1}\right)=\left\langle\theta^{1}, \mu^{1}\right\rangle-A\left(\theta^{1}\right)$ 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

$$
A^{*}(\mu)=\sup _{\theta \in \Omega}(\langle\theta, \mu\rangle-A(\theta)) \geq\left\langle\theta^{\prime}, \mu\right\rangle-A\left(\theta^{\prime}\right) \text { for any } \theta^{\prime} \in \Omega
$$

- We can also write the KL as:

$$
\begin{align*}
D\left(\theta^{1} \| \theta^{2}\right) & =A\left(\theta^{2}\right)-A\left(\theta^{1}\right)-\left\langle\mu^{1}, \theta^{2}-\theta^{1}\right\rangle  \tag{17.34}\\
& =A\left(\theta^{2}\right)-\left\langle\mu^{1}, \theta^{2}\right\rangle-\left[A\left(\theta^{1}\right)-\left\langle\mu^{1}, \theta^{1}\right\rangle\right]  \tag{17.35}\\
& =A\left(\theta^{2}\right)-\left\langle\mu^{1}, \theta^{2}\right\rangle+A^{*}\left(\mu^{1}\right) \triangleq D\left(\mu^{1} \| \theta^{2}\right) \tag{17.36}
\end{align*}
$$

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

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

$$
\begin{equation*}
A(\theta)=\sup _{\mu \in \mathcal{M}}\left\{\langle\theta, \mu\rangle-A^{*}(\mu)\right\} \tag{13.7}
\end{equation*}
$$

## Mean field, KL-Divergence, Exponential Model Families

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

$$
\begin{align*}
D\left(\theta^{1} \| \theta^{2}\right) & =A\left(\theta^{2}\right)-A\left(\theta^{1}\right)-\left\langle\mu^{1}, \theta^{2}-\theta^{1}\right\rangle  \tag{17.34}\\
& =A\left(\theta^{2}\right)-\left\langle\mu^{1}, \theta^{2}\right\rangle-\left[A\left(\theta^{1}\right)-\left\langle\mu^{1}, \theta^{1}\right\rangle\right]  \tag{17.35}\\
& \left.=A\left(\theta^{2}\right)-\left\langle\mu^{1}, \theta^{2}\right\rangle+A^{*}\left(\mu^{1}\right)\right\rangle  \tag{17.36}\\
& \left.\Delta \mu^{1}| | \theta^{2}\right)
\end{align*}
$$

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

- In particular, this equation (variational expression for the cumulant):
- . . . can tre written as:

$$
\begin{equation*}
\inf _{\mu \in \mathcal{M}}\left\{A(\theta)+A^{*}(\mu)-\langle\theta, \mu\rangle\right\} \neq \inf _{\mu \in \mathcal{M}} D(\mu \| \theta)=0 \tag{17.37}
\end{equation*}
$$

## Mean field, KL-Divergence, Exponential Model Families

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

$$
\begin{equation*}
\max _{\mu \in \mathcal{M}_{F}(G)}\left\{\langle\mu, \theta\rangle-A_{F}^{*}(\mu)\right\}=\max _{\mu \in \mathcal{M}_{F}(G)}\left\{\langle\mu, \theta\rangle-A^{*}(\mu)\right\} \tag{17.38}
\end{equation*}
$$

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:

$$
\begin{equation*}
\max _{\mu \in \mathcal{M}_{F}(G)}\left\{\langle\mu, \theta\rangle-A_{F}^{*}(\mu)\right\}=\max _{\mu \in \mathcal{M}_{F}(G)}\left\{\langle\mu, \theta\rangle-A^{*}(\mu)\right\} \tag{17.38}
\end{equation*}
$$

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 $p_{\theta}$, 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: $\mu_{s}=\mathbb{E}\left[X_{s}\right]=p\left(X_{s}=1\right)$, $\mu_{s t}=\mathbb{E}\left[X_{s} X_{t}\right]=p\left(X_{s}=1, X_{t}=1\right)$, 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: $\mu_{s}=\mathbb{E}\left[X_{s}\right]=p\left(X_{s}=1\right)$, $\mu_{s t}=\mathbb{E}\left[X_{s} X_{t}\right]=p\left(X_{s}=1, X_{t}=1\right)$, thus $\mu \in \mathbb{R}^{|V|+|E|}$.
- Let $F_{0}=(V, \emptyset)$ be our mean field approximation family. Thus,

$$
\mathcal{M}_{F_{0}}(G)=\left\{\mu \in \mathbb{R}^{|V|+|E|} \mid 0 \leq \mu_{s} \leq 1 \quad \forall s \in V, \text { and } \mu_{s t}=\mu_{s} \mu_{t} \forall\right\}
$$

## Naïve Mean field for Ising Model

- A classic example of mean-field (goes back to statistical physics)
- Mean parameters for Ising: $\mu_{s}=\mathbb{E}\left[X_{s}\right]=p\left(X_{s}=1\right)$, $\mu_{s t}=\mathbb{E}\left[X_{s} X_{t}\right]=p\left(X_{s}=1, X_{t}=1\right)$, thus $\mu \in \mathbb{R}^{|V|+|E|}$.
- Let $F_{0}=(V, \emptyset)$ be our mean field approximation family. Thus,

$$
\mathcal{M}_{F_{0}}(G)=\left\{\mu \in \mathbb{R}^{|V|+|E|} \mid 0 \leq \mu_{s} \leq 1 \quad \forall s \in V, \text { and } \mu_{s t}=\mu_{s} \mu_{t} \forall\right\}
$$

- Key is that for $\mu \in \mathcal{M}_{F_{0}}(G)$, dual is not hard to calculate, that is

$$
\begin{equation*}
-A_{F_{0}}^{*}(\mu)=\sum_{s \in V} H_{s}\left(\mu_{s}\right) \tag{17.39}
\end{equation*}
$$

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: $\mu_{s}=\mathbb{E}\left[X_{s}\right]=p\left(X_{s}=1\right)$, $\mu_{s t}=\mathbb{E}\left[X_{s} X_{t}\right]=p\left(X_{s}=1, X_{t}=1\right)$, thus $\mu \in \mathbb{R}^{|V|+|E|}$.
- Let $F_{0}=(V, \emptyset)$ be our mean field approximation family. Thus,

$$
\mathcal{M}_{F_{0}}(G)=\left\{\mu \in \mathbb{R}^{|V|+|E|} \mid 0 \leq \mu_{s} \leq 1 \quad \forall s \in V, \text { and } \mu_{s t}=\mu_{s} \mu_{t} \forall\right\}
$$

- Key is that for $\mu \in \mathcal{M}_{F_{0}}(G)$, dual is not hard to calculate, that is

$$
\begin{equation*}
-A_{F_{0}}^{*}(\mu)=\sum_{s \in V} H_{s}\left(\mu_{s}\right) \tag{17.39}
\end{equation*}
$$

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



## Naive Mean field for Ising Model

- We get variational lower bound problem

$$
A(\theta) \geq \max _{\left(\mu_{1}, \ldots, \mu_{m}\right) \in[0,1]^{m}}\left\{\sum_{s \in V} \theta_{s} \mu_{s}+\sum_{(s, t) \in E} \theta_{s t} \mu_{s} \mu_{t}+\sum_{s \in V} H_{s}\left(\mu_{s}\right)\right\}
$$

- Have constrained form of edge mean parameters $\mu_{s t}=\mu_{s} \mu_{t}$


## Naive Mean field for Ising Model

- We get variational lower bound problem

$$
A(\theta) \geq \max _{\left(\mu_{1}, \ldots, \mu_{m}\right) \in[0,1]^{m}}\left\{\sum_{s \in V} \theta_{s} \mu_{s}+\sum_{(s, t) \in E} \theta_{s t} \mu_{s} \mu_{t}+\sum_{s \in V} H_{s}\left(\mu_{s}\right)\right\}
$$

- Have constrained form of edge mean parameters $\mu_{s t}=\mu_{s} \mu_{t}$
- $\left(\mu_{1}, \ldots, \mu_{m}\right) \in[0,1]^{m}$ is $m$-D hypercube.


## Naive Mean field for Ising Model

- We get variational lower bound problem

$$
A(\theta) \geq \max _{\left(\mu_{1}, \ldots, \mu_{m}\right) \in[0,1]^{m}}\left\{\sum_{s \in V} \theta_{s} \mu_{s}+\sum_{(s, t) \in E} \theta_{s t} \mu_{s} \mu_{t}+\sum_{s \in V} H_{s}\left(\mu_{s}\right)\right\}
$$

(17.40)

- Have constrained form of edge mean parameters $\mu_{s t}=\mu_{s} \mu_{t}$
- $\left(\mu_{1}, \ldots, \mu_{m}\right) \in[0,1]^{m}$ is $m$-D hypercube.
- Once again, we have a non-convex problem.


## Naive Mean field for Ising Model

- We get variational lower bound problem

$$
A(\theta) \geq \max _{\left(\mu_{1}, \ldots, \mu_{m}\right) \in[0,1]^{m}}\left\{\sum_{s \in V} \theta_{s} \mu_{s}+\sum_{(s, t) \in E} \theta_{s t} \mu_{s} \mu_{t}+\sum_{s \in V} H_{s}\left(\mu_{s}\right)\right\}
$$

- Have constrained form of edge mean parameters $\mu_{s t}=\mu_{s} \mu_{t}$
- $\left(\mu_{1}, \ldots, \mu_{m}\right) \in[0,1]^{m}$ 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

$$
\begin{equation*}
A(\theta) \geq \max _{\left(\mu_{1}, \ldots, \mu_{m}\right) \in[0,1]^{m}}\left\{\sum_{s \in V} \theta_{s} \mu_{s}+\sum_{(s, t) \in E} \theta_{s t} \mu_{s} \mu_{t}+\sum_{s \in V} H_{s}\left(\mu_{s}\right)\right\} \tag{17.40}
\end{equation*}
$$

- Have constrained form of edge mean parameters $\mu_{s t}=\mu_{s} \mu_{t}$
- $\left(\mu_{1}, \ldots, \mu_{m}\right) \in[0,1]^{m}$ 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

$$
\begin{equation*}
A(\theta) \geq \max _{\left(\mu_{1}, \ldots, \mu_{m}\right) \in[0,1]^{m}}\left\{\sum_{s \in V} \theta_{s} \mu_{s}+\sum_{(s, t) \in E} \theta_{s t} \mu_{s} \mu_{t}+\sum_{s \in V} H_{s}\left(\mu_{s}\right)\right\} \tag{17.40}
\end{equation*}
$$

- Have constrained form of edge mean parameters $\mu_{s t}=\mu_{s} \mu_{t}$
- $\left(\mu_{1}, \ldots, \mu_{m}\right) \in[0,1]^{m}$ 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 $\mu_{s}$ fixing all $\mu_{t}$ for $t \neq s$.


## Naive Mean field for Ising Model

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

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 $\mu_{s}$ fixing all $\mu_{t}$ for $t \neq s$.
- Taking derivatives w.r.t. $\mu_{s}$, we get the following update rule for element $\mu_{s}$

$$
\begin{equation*}
\mu_{s} \leftarrow \sigma\left(\theta_{s}+\sum_{t \in N(s)} \theta_{s t} \mu_{t}\right) \tag{17.41}
\end{equation*}
$$

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 $\mu_{s}$ fixing all $\mu_{t}$ for $t \neq s$.
- Taking derivatives w.r.t. $\mu_{s}$, we get the following update rule for element $\mu_{s}$

$$
\begin{equation*}
\mu_{s} \leftarrow \sigma\left(\theta_{s}+\sum_{t \in N(s)} \theta_{s t} \mu_{t}\right) \tag{17.41}
\end{equation*}
$$

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 $\left(X_{1}, X_{2}\right), X_{i} \in\{-1,+1\}$.


## Example of Lack of Convexity

- Consider simple two variable example $\left(X_{1}, X_{2}\right), X_{i} \in\{-1,+1\}$.
- Exponential family form

$$
\begin{equation*}
p_{\theta}(x) \propto \exp \left(\theta_{1} x_{1}+\theta_{2} x_{2}+\theta_{12} x_{1} x_{2}\right) \tag{17.42}
\end{equation*}
$$

having mean parameters $\mu_{i}=\mathbb{E}\left[X_{i}\right]$ and $\mu_{12}=\mathbb{E}\left[X_{1} X_{2}\right]$.

## Example of Lack of Convexity

- Consider simple two variable example $\left(X_{1}, X_{2}\right), X_{i} \in\{-1,+1\}$.
- Exponential family form

$$
\begin{equation*}
p_{\theta}(x) \propto \exp \left(\theta_{1} x_{1}+\theta_{2} x_{2}+\theta_{12} x_{1} x_{2}\right) \tag{17.42}
\end{equation*}
$$

having mean parameters $\mu_{i}=\mathbb{E}\left[X_{i}\right]$ and $\mu_{12}=\mathbb{E}\left[X_{1} X_{2}\right]$.

- Impose constraint $\mu_{12}=\mu_{1} \mu_{2}$, we get mean field objective

$$
\begin{equation*}
f\left(\mu_{1}, \mu_{2} ; \theta\right)=\theta_{12} \mu_{1} \mu_{2}+\theta_{1} \mu_{1}+\theta_{2} \mu_{2}+H\left(\mu_{1}\right)+H\left(\mu_{2}\right) \tag{17.43}
\end{equation*}
$$

where $H\left(\mu_{i}\right)=-\frac{1}{2}\left(1+\mu_{i}\right) \log \frac{1}{2}\left(1+\mu_{i}\right)-\frac{1}{2}\left(1-\mu_{i}\right) \log \frac{1}{2}\left(1-\mu_{i}\right)$
Note that $p\left(X_{i}=+1\right)=\frac{1}{2}\left(1+\mu_{i}\right)$

## Example of Lack of Convexity

- Consider simple two variable example $\left(X_{1}, X_{2}\right), X_{i} \in\{-1,+1\}$.
- Exponential family form

$$
\begin{equation*}
p_{\theta}(x) \propto \exp \left(\theta_{1} x_{1}+\theta_{2} x_{2}+\theta_{12} x_{1} x_{2}\right) \tag{17.42}
\end{equation*}
$$

having mean parameters $\mu_{i}=\mathbb{E}\left[X_{i}\right]$ and $\mu_{12}=\mathbb{E}\left[X_{1} X_{2}\right]$.

- Impose constraint $\mu_{12}=\mu_{1} \mu_{2}$, we get mean field objective

$$
\begin{equation*}
f\left(\mu_{1}, \mu_{2} ; \theta\right)=\theta_{12} \mu_{1} \mu_{2}+\theta_{1} \mu_{1}+\theta_{2} \mu_{2}+H\left(\mu_{1}\right)+H\left(\mu_{2}\right) \tag{17.43}
\end{equation*}
$$

where $H\left(\mu_{i}\right)=-\frac{1}{2}\left(1+\mu_{i}\right) \log \frac{1}{2}\left(1+\mu_{i}\right)-\frac{1}{2}\left(1-\mu_{i}\right) \log \frac{1}{2}\left(1-\mu_{i}\right)$

- Consider sub-models of the form:

$$
\begin{equation*}
\left(\theta_{1}, \theta_{2}, \theta_{12}\right)=\left(0,0, \frac{1}{4} \log \frac{q}{1-q}\right) \triangleq \theta(q) \tag{17.44}
\end{equation*}
$$

where $q \in(0,1)$ is a parameter such that, for any $q$ we have $\mathbb{E}\left[X_{i}\right]=0$. It turns out that in this form, we have $q=p\left(X_{1}=X_{2}\right)$.

## Example of Lack of Convexity

- Consider simple two variable example $\left(X_{1}, X_{2}\right), X_{i} \in\{-1,+1\}$.
- Exponential family form

$$
\begin{equation*}
p_{\theta}(x) \propto \exp \left(\theta_{1} x_{1}+\theta_{2} x_{2}+\theta_{12} x_{1} x_{2}\right) \tag{17.42}
\end{equation*}
$$

having mean parameters $\mu_{i}=\mathbb{E}\left[X_{i}\right]$ and $\mu_{12}=\mathbb{E}\left[X_{1} X_{2}\right]$.

- Impose constraint $\mu_{12}=\mu_{1} \mu_{2}$, we get mean field objective

$$
\begin{equation*}
f\left(\mu_{1}, \mu_{2} ; \theta\right)=\theta_{12} \mu_{1} \mu_{2}+\theta_{1} \mu_{1}+\theta_{2} \mu_{2}+H\left(\mu_{1}\right)+H\left(\mu_{2}\right) \tag{17.43}
\end{equation*}
$$

where $H\left(\mu_{i}\right)=-\frac{1}{2}\left(1+\mu_{i}\right) \log \frac{1}{2}\left(1+\mu_{i}\right)-\frac{1}{2}\left(1-\mu_{i}\right) \log \frac{1}{2}\left(1-\mu_{i}\right)$

- Consider sub-models of the form:

$$
\begin{equation*}
\left(\theta_{1}, \theta_{2}, \theta_{12}\right)=\left(0,0, \frac{1}{4} \log \frac{q}{1-q}\right) \triangleq \theta(q) \tag{17.44}
\end{equation*}
$$

where $q \in(0,1)$ is a parameter such that, for any $q$ we have $\mathbb{E}\left[X_{i}\right]=0$. It turns out that in this form, we have $q=p\left(X_{1}=X_{2}\right)$.

- Is mean field objective in this case convex for all $q$ ?


## Lack of Convexity example

- For $q=0.5$, objective $f\left(\mu_{1}, \mu_{2} ; \theta(0.5)\right)$ has global maximum at $\left(\mu_{1}, \mu_{2}\right)=(0,0)$ so mean field is exact and convex. This corresponds to $p\left(X_{1}=X_{2}\right)=0$.


## Lack of Convexity example

- For $q=0.5$, objective $f\left(\mu_{1}, \mu_{2} ; \theta(0.5)\right)$ has global maximum at $\left(\mu_{1}, \mu_{2}\right)=(0,0)$ so mean field is exact and convex. This corresponds to $p\left(X_{1}=X_{2}\right)=0$.
- When $q$ gets small, $f$ becomes non-convex, e.g., has multiple modes in figure.


## Lack of Convexity example

- For $q=0.5$, objective $f\left(\mu_{1}, \mu_{2} ; \theta(0.5)\right)$ has global maximum at $\left(\mu_{1}, \mu_{2}\right)=(0,0)$ so mean field is exact and convex. This corresponds to $p\left(X_{1}=X_{2}\right)=0$.
- 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 $\mu(F)=\left(\mu_{\alpha}, \alpha \in \mathcal{I}(F)\right)$.


## Structured Mean Field

- Again, $\mathcal{I}(F)$ is set of suff. stats. corresponding to $F$, and we have corresponding mean vector $\mu(F)=\left(\mu_{\alpha}, \alpha \in \mathcal{I}(F)\right)$.
- 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 $\mu(F)=\left(\mu_{\alpha}, \alpha \in \mathcal{I}(F)\right)$.
- 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 $\mu(F)=\left(\mu_{\alpha}, \alpha \in \mathcal{I}(F)\right)$.
- 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 $\mu$ (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 $\mu(F)=\left(\mu_{\alpha}, \alpha \in \mathcal{I}(F)\right)$.
- 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 $\mu$ (the other values are derivations from entries within $\mu(F)$.
- Other mean parameters $\mu_{\beta}$ for $\beta \in \mathcal{I} \backslash \mathcal{I}(F)$ do play a role in the value of the mean field variational problem but their value is derivable from values $\mu(F)$, thus we can express the $\mu_{\beta}$ in functional form based on values $\mu(F)$.


## Structured Mean Field

- Again, $\mathcal{I}(F)$ is set of suff. stats. corresponding to $F$, and we have corresponding mean vector $\mu(F)=\left(\mu_{\alpha}, \alpha \in \mathcal{I}(F)\right)$.
- 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 $\mu$ (the other values are derivations from entries within $\mu(F)$.
- Other mean parameters $\mu_{\beta}$ for $\beta \in \mathcal{I} \backslash \mathcal{I}(F)$ do play a role in the value of the mean field variational problem but their value is derivable from values $\mu(F)$, thus we can express the $\mu_{\beta}$ in functional form based on values $\mu(F)$.
- Thus, for each $\beta \in \mathcal{I} \backslash \mathcal{I}(F)$, we set $\mu_{\beta}=g_{\beta}(\mu(F))$ for function $g_{\beta}$.


## Structured Mean Field

- Again, $\mathcal{I}(F)$ is set of suff. stats. corresponding to $F$, and we have corresponding mean vector $\mu(F)=\left(\mu_{\alpha}, \alpha \in \mathcal{I}(F)\right)$.
- 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 $\mu$ (the other values are derivations from entries within $\mu(F)$.
- Other mean parameters $\mu_{\beta}$ for $\beta \in \mathcal{I} \backslash \mathcal{I}(F)$ do play a role in the value of the mean field variational problem but their value is derivable from values $\mu(F)$, thus we can express the $\mu_{\beta}$ in functional form based on values $\mu(F)$.
- Thus, for each $\beta \in \mathcal{I} \backslash \mathcal{I}(F)$, we set $\mu_{\beta}=g_{\beta}(\mu(F))$ for function $g_{\beta}$.
- Example: mean field Ising, $\mu_{s t}=g(\mu(F))=\mu_{s} \mu_{t}$.


## Structured Mean Field

- The mean field optimization problem becomes

$$
\begin{aligned}
& \max _{\mu \in \mathcal{M}_{F}(G)}\left\{\langle\mu, \theta\rangle-A_{F}^{*}(\mu)\right\} \\
& =\max _{\mu(F) \in \mathcal{M}(F)}\{\underbrace{\left.\sum_{\alpha \in \mathcal{I}(F)} \theta_{\alpha} \mu_{\alpha}+\sum_{\alpha \in \mathcal{I}^{c}(F)} \theta_{\alpha} g_{\alpha}(\mu(F))-A_{F}^{*}(\mu(F))\right\}}_{f(\mu(F))}
\end{aligned}
$$

(17.46)

## Structured Mean Field

- The mean field optimization problem becomes

$$
\begin{align*}
\max _{\mu \in \mathcal{M}_{F}(G)}\left\{\langle\mu, \theta\rangle-A_{F}^{*}(\mu)\right\}  \tag{17.45}\\
=\max _{\mu(F) \in \mathcal{M}(F)}\{\underbrace{\sum_{\alpha \in \mathcal{I}(F)} \theta_{\alpha} \mu_{\alpha}+\sum_{\alpha \in \mathcal{I}^{c}(F)} \theta_{\alpha} g_{\alpha}(\mu(F))-A_{F}^{*}(\mu(F))}_{f(\mu(F))}\}
\end{align*}
$$

(17.46)

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

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

## Structured Mean Field

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

$$
\begin{equation*}
\nabla A_{F}^{*}(\mu(F))=\theta+\sum_{\alpha \in \mathcal{I}(G) \backslash \mathcal{I}(F)} \theta_{\alpha} \nabla g_{\alpha}(\mu(F)) \tag{17.48}
\end{equation*}
$$

## Structured Mean Field

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

$$
\begin{equation*}
\nabla A_{F}^{*}(\mu(F))=\theta+\sum_{\alpha \in \mathcal{I}(G) \backslash \mathcal{I}(F)} \theta_{\alpha} \nabla g_{\alpha}(\mu(F)) \tag{17.48}
\end{equation*}
$$

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

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

## Structured Mean Field

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

$$
\begin{equation*}
\nabla A_{F}^{*}(\mu(F))=\theta+\sum_{\alpha \in \mathcal{I}(G) \backslash \mathcal{I}(F)} \theta_{\alpha} \nabla g_{\alpha}(\mu(F)) \tag{17.48}
\end{equation*}
$$

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

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

- Above is the mean field update, mapping from a canonical parameters $\left(\theta_{\beta}\right.$ for $\left.\beta \in \mathcal{I}(F)\right)$ and using the mean parameters $\mu(F)$ to new updated canonical parameters $\gamma_{\beta}(F)$ for $\left.\beta \in \mathcal{I}(F)\right)$. 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 $\nabla A$ is the forward mapping, maping from mean to canonical.


## Structured Mean Field

- Alternatively, we can transform back to mean parameters right away using $\nabla A$ 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 $\nabla A$ 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 $\beta \in \mathcal{I}(F)$ :

$$
\begin{equation*}
\mu_{\beta}(F) \leftarrow \frac{\partial A_{F}}{\partial \gamma_{\beta}}\left(\theta_{\beta}+\sum_{\alpha \in \mathcal{I}(G) \backslash \mathcal{I}(F)} \theta_{\alpha} \nabla g_{\alpha}(\mu(F))\right) \tag{17.50}
\end{equation*}
$$

## Structured Mean Field

- Alternatively, we can transform back to mean parameters right away using $\nabla A$ 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 $\beta \in \mathcal{I}(F)$ :

$$
\begin{equation*}
\mu_{\beta}(F) \leftarrow \frac{\partial A_{F}}{\partial \gamma_{\beta}}\left(\theta_{\beta}+\sum_{\alpha \in \mathcal{I}(G) \backslash \mathcal{I}(F)} \theta_{\alpha} \nabla g_{\alpha}(\mu(F))\right) \tag{17.50}
\end{equation*}
$$

- 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 $y_{1: T}^{i}$ for $i \in[M]$, coupled together at each time via factor $p\left(\bar{y}_{t} \mid x_{t}^{1}, x_{t}^{2}, \ldots, x_{t}^{M}\right)$.


## Structured Mean Field Factorial HMMs

- This idea was developed and applied using factorial HMMs.

- Graph consists of $M$ 1st-order Markov chains $y_{1: T}^{i}$ for $i \in[M]$, coupled together at each time via factor $p\left(\bar{y}_{t} \mid x_{t}^{1}, x_{t}^{2}, \ldots, x_{t}^{M}\right)$.
- 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 $y_{1: T}^{i}$ for $i \in[M]$, coupled together at each time via factor $p\left(\bar{y}_{t} \mid x_{t}^{1}, x_{t}^{2}, \ldots, x_{t}^{M}\right)$.
- 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 $F=\left\{\left\{x_{t}^{i}, x_{t+1}^{i}\right\}: i \in[M], t \in[T]\right\}$ and remaining structure $E \backslash F=\left\{\left\{x_{t}^{1}, x_{t}^{2}, \ldots, x_{t}^{M}\right\}: t \in[T]\right\}$.


## 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 $\beta \in \mathcal{I} \backslash \mathcal{I}(F)$, derivable functions have form $g_{\beta}(\mu(F))=\prod_{i=1}^{M} f_{i}\left(\left\{\mu_{i}(F)\right\}\right)$, 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 $\beta \in \mathcal{I} \backslash \mathcal{I}(F)$, derivable functions have form $g_{\beta}(\mu(F))=\prod_{i=1}^{M} f_{i}\left(\left\{\mu_{i}(F)\right\}\right)$, 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 $\beta \in \mathcal{I} \backslash \mathcal{I}(F)$, derivable functions have form $g_{\beta}(\mu(F))=\prod_{i=1}^{M} f_{i}\left(\left\{\mu_{i}(F)\right\}\right)$, 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, $O\left(M T r^{2}\right)$.


## 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 $H_{\text {app }}=\sum_{g \in E} c(g) H_{g}\left(\tau_{g}\right)$ (via Möbius) to get Kikuchi variational approximation, message passing on hypergraphs.
(3) Partition $\tau$ into $(\tau, \tilde{\tau})$, and set $\mathcal{M} \leftarrow \mathcal{L}(\phi, \Phi)$ and set $-A^{*}(\mu) \leftarrow H_{\text {ep }}(\tau, \tilde{\tau})$ to get expectation propagation.
(9) Mean field (from variational perspective) is (with $\left.\mathcal{M}_{F}(G) \subseteq \mathcal{M}\right)$

$$
\begin{equation*}
A(\theta) \geq \max _{\mu \in \mathcal{M}_{F}(G)}\left\{\langle\mu, \theta\rangle-A_{F}^{*}(\mu)\right\}=A_{\mathrm{mf}}(\theta) \tag{17.1}
\end{equation*}
$$

## Convex Relaxations and Upper Bounds

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

- What about upper bounds?


## Convex Relaxations and Upper Bounds

$$
\begin{equation*}
A(\theta)=\sup _{\mu \in \mathcal{M}}\left\{\langle\theta, \mu\rangle-A^{*}(\mu)\right\} \tag{17.51}
\end{equation*}
$$

- 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

$$
\begin{equation*}
A(\theta)=\sup _{\mu \in \mathcal{M}}\left\{\langle\theta, \mu\rangle-A^{*}(\mu)\right\} \tag{17.51}
\end{equation*}
$$

- 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

$$
\begin{equation*}
A(\theta)=\sup _{\mu \in \mathcal{M}}\left\{\langle\theta, \mu\rangle-A^{*}(\mu)\right\} \tag{17.51}
\end{equation*}
$$

- 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

$$
\begin{equation*}
A(\theta)=\sup _{\mu \in \mathcal{M}}\left\{\langle\theta, \mu\rangle-A^{*}(\mu)\right\} \tag{17.51}
\end{equation*}
$$

- 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 $\phi=\left(\phi_{\alpha}, \alpha \in \mathcal{I}\right)$ and canonical parameters $\theta=\left(\theta_{\alpha}, \alpha \in \mathcal{I}\right)$.


## Convex Relaxations and Upper Bounds - Relaxed Entropy

- Recall sufficient stats $\phi=\left(\phi_{\alpha}, \alpha \in \mathcal{I}\right)$ and canonical parameters $\theta=\left(\theta_{\alpha}, \alpha \in \mathcal{I}\right)$.
- In general, inference (computing mean parameters) is hard for a given $G$.


## Convex Relaxations and Upper Bounds - Relaxed Entropy

- Recall sufficient stats $\phi=\left(\phi_{\alpha}, \alpha \in \mathcal{I}\right)$ and canonical parameters $\theta=\left(\theta_{\alpha}, \alpha \in \mathcal{I}\right)$.
- 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 $\phi=\left(\phi_{\alpha}, \alpha \in \mathcal{I}\right)$ and canonical parameters $\theta=\left(\theta_{\alpha}, \alpha \in \mathcal{I}\right)$.
- 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 $\phi=\left(\phi_{\alpha}, \alpha \in \mathcal{I}\right)$ and canonical parameters $\theta=\left(\theta_{\alpha}, \alpha \in \mathcal{I}\right)$.
- 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.
- l.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 $\phi=\left(\phi_{\alpha}, \alpha \in \mathcal{I}\right)$ and canonical parameters $\theta=\left(\theta_{\alpha}, \alpha \in \mathcal{I}\right)$.
- 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.
- l.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 $\phi=\left(\phi_{\alpha}, \alpha \in \mathcal{I}\right)$ and canonical parameters $\theta=\left(\theta_{\alpha}, \alpha \in \mathcal{I}\right)$.
- 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.
- l.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

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

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


## 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 $H(\mu(F)) \triangleq H\left(p_{\mu(F)}\right)=-A^{*}(\mu(F))$.
- Critically, we have that $H(\mu(F)) \geq H(\mu)=H\left(p_{\mu}\right)$, 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

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

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

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


## 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

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

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

- Intuition: $H(\mu)=H\left(p_{\mu}\right)$ is the entropy of the exponential family model with mean parameters $\mu$.
- equivalently $H(\mu)=H\left(p_{\mu}\right)$ 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

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

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

- Intuition: $H(\mu)=H\left(p_{\mu}\right)$ is the entropy of the exponential family model with mean parameters $\mu$.
- equivalently $H(\mu)=H\left(p_{\mu}\right)$ 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

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

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

- Intuition: $H(\mu)=H\left(p_{\mu}\right)$ is the entropy of the exponential family model with mean parameters $\mu$.
- equivalently $H(\mu)=H\left(p_{\mu}\right)$ 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

$$
\begin{equation*}
A^{*}(\mu)=\sup _{\theta \in \mathbb{R}^{d}}\{\langle\mu, \theta\rangle-A(\theta)\} \tag{17.54}
\end{equation*}
$$

## Convex Relaxations and Upper Bounds - Relaxed Entropy

Proof.

- Dual problem

$$
\begin{equation*}
A^{*}(\mu)=\sup _{\theta \in \mathbb{R}^{d}}\{\langle\mu, \theta\rangle-A(\theta)\} \tag{17.54}
\end{equation*}
$$

- Dual problem in sub-graph case.

$$
\begin{equation*}
A^{*}(\mu(F))=\sup _{\theta(F) \in \mathbb{R}^{d(F)}}\{\langle\mu(F), \theta(F)\rangle-A(\theta(F))\} \tag{17.55}
\end{equation*}
$$

## Convex Relaxations and Upper Bounds - Relaxed Entropy

## Proof.

- Dual problem

$$
\begin{equation*}
A^{*}(\mu)=\sup _{\theta \in \mathbb{R}^{d}}\{\langle\mu, \theta\rangle-A(\theta)\} \tag{17.54}
\end{equation*}
$$

- Dual problem in sub-graph case.

$$
\begin{equation*}
A^{*}(\mu(F))=\sup _{\theta(F) \in \mathbb{R}^{d(F)}}\{\langle\mu(F), \theta(F)\rangle-A(\theta(F))\} \tag{17.55}
\end{equation*}
$$

- Dual problem in sub-graph case - alternate expression

$$
\begin{gathered}
A^{*}(\mu(F))=\sup _{\substack{\theta \in \mathbb{R}^{d}}}\{\langle\mu, \theta\rangle-A(\theta)\} \\
\theta_{\alpha}=0 \forall \alpha \notin \mathcal{I}(F)
\end{gathered}
$$

## Convex Relaxations and Upper Bounds - Relaxed Entropy

## Proof.

- Dual problem

$$
\begin{equation*}
A^{*}(\mu)=\sup _{\theta \in \mathbb{R}^{d}}\{\langle\mu, \theta\rangle-A(\theta)\} \tag{17.54}
\end{equation*}
$$

- Dual problem in sub-graph case.

$$
\begin{equation*}
A^{*}(\mu(F))=\sup _{\theta(F) \in \mathbb{R}^{d(F)}}\{\langle\mu(F), \theta(F)\rangle-A(\theta(F))\} \tag{17.55}
\end{equation*}
$$

- Dual problem in sub-graph case - alternate expression

$$
\begin{gathered}
A^{*}(\mu(F))=\sup _{\substack{\theta \in \mathbb{R}^{d}}}\{\langle\mu, \theta\rangle-A(\theta)\} \\
\theta_{\alpha}=0 \forall \alpha \notin \mathcal{I}(F)
\end{gathered}
$$

- 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

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

## 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

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

- 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

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

- 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:

$$
\begin{align*}
\mathcal{L}(G ; \mathfrak{D}) & =\left\{\tau \in \mathbb{R}^{d} \mid \tau(F) \in \mathcal{M}(F) \forall F \in \mathfrak{D}\right\}  \tag{17.58}\\
& =\bigcap_{F \in \mathfrak{O}} \mathcal{M}(F) \tag{17.59}
\end{align*}
$$

## 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:

$$
\begin{align*}
\mathcal{L}(G ; \mathfrak{D}) & =\left\{\tau \in \mathbb{R}^{d} \mid \tau(F) \in \mathcal{M}(F) \forall F \in \mathfrak{D}\right\}  \tag{17.58}\\
& =\bigcap_{F \in \mathfrak{D}} \mathcal{M}(F) \tag{17.59}
\end{align*}
$$

- 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:

$$
\begin{align*}
\mathcal{L}(G ; \mathfrak{D}) & =\left\{\tau \in \mathbb{R}^{d} \mid \tau(F) \in \mathcal{M}(F) \forall F \in \mathfrak{D}\right\}  \tag{17.58}\\
& =\bigcap_{F \in \mathfrak{D}} \mathcal{M}(F) \tag{17.59}
\end{align*}
$$

- 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.

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

## 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.

$$
\begin{equation*}
B_{\mathfrak{D}}(\theta ; \rho) \triangleq \sup _{\tau \in \mathcal{L}(G ; \mathfrak{Q})}\left\{\langle\tau, \theta\rangle+\sum_{F \in \mathfrak{D}} \rho(F) H(\tau(F))\right\} \tag{17.60}
\end{equation*}
$$

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


## 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.

$$
\begin{equation*}
B_{\mathfrak{D}}(\theta ; \rho) \triangleq \sup _{\tau \in \mathcal{L}(G ; \mathfrak{Q})}\left\{\langle\tau, \theta\rangle+\sum_{F \in \mathfrak{D}} \rho(F) H(\tau(F))\right\} \tag{17.60}
\end{equation*}
$$

- Objective is convex in $\theta$ since it is a max over a set of affine functions of $\theta$ (i.e., $g(\theta)=\max _{\tau}\langle\tau, \theta\rangle+c_{\tau}$ )
- Also, $\mathcal{L}(G ; \mathfrak{D})$ 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.

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

- Objective is convex in $\theta$ since it is a max over a set of affine functions of $\theta$ (i.e., $g(\theta)=\max _{\tau}\langle\tau, \theta\rangle+c_{\tau}$ )
- Also, $\mathcal{L}(G ; \mathfrak{D})$ 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

