Probabilistic Graphical Models

Brown University CSCI 2950-P, Spring 2013
Prof. Erik Sudderth

Lecture 22:
Reparameterization & Loopy BP,
Reweighted Belief Propagation

Some figures and examples courtesy M. Wainwright & M. Jordan,
Graphical Models, Exponential Families, & Variational Inference, 2008.

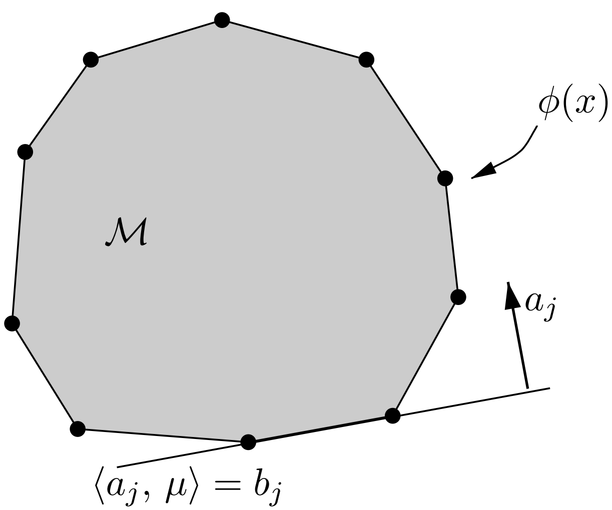
Discrete Variables & Marginal Polytopes

$$p(x \mid \theta) = \exp\{\theta^T \phi(x) - A(\theta)\} \qquad A(\theta) = \log \sum_{\mathcal{X}} \exp\{\theta^T \phi(x)\}$$

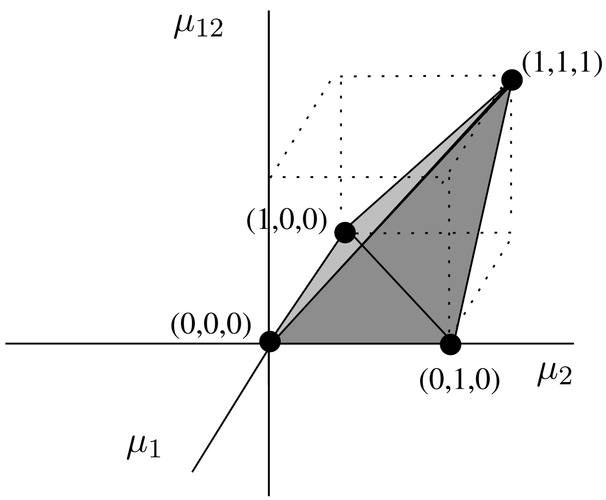
$$\mu = \nabla_\theta A(\theta) = \mathbb{E}_\theta[\phi(x)] = \sum_{\mathcal{X}} \phi(x) p(x \mid \theta)$$

$$\mathcal{M} \triangleq \{\mu \in \mathbb{R}^d \mid \exists\, p \text{ such that } \mathbb{E}_p[\phi(x)] = \mu\} \subseteq [0,1]^d$$

$$\mathcal{M} = \mathrm{conv}\{\phi(x) \mid x \in \mathcal{X}\}$$ *convex hull of possible configurations*



General Convex Polytope

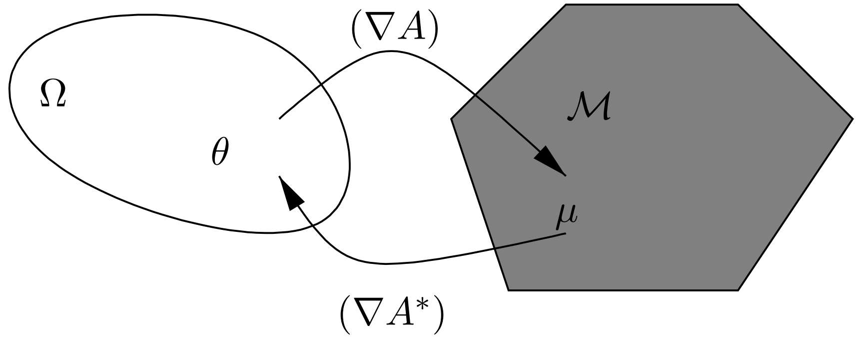


Pair of Binary Variables

Inference as Optimization

$$p(x \mid \theta) = \exp\{\theta^T \phi(x) - A(\theta)\}$$

$$A(\theta) = \log \sum_{x \in \mathcal{X}} \exp\{\theta^T \phi(x)\}$$



- Express log-partition as optimization over all distributions $\mathcal{Q}$

$$A(\theta) = \sup_{q \in \mathcal{Q}} \left\{ \sum_{x \in \mathcal{X}} \theta^T \phi(x) q(x) - \sum_{x \in \mathcal{X}} q(x) \log q(x) \right\}$$

Jensen's inequality gives arg max: $q(x) = p(x \mid \theta)$

- More compact to optimize over relevant *sufficient statistics*:

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \left\{ \theta^T \mu + H(p(x \mid \theta(\mu)) \right\}$$

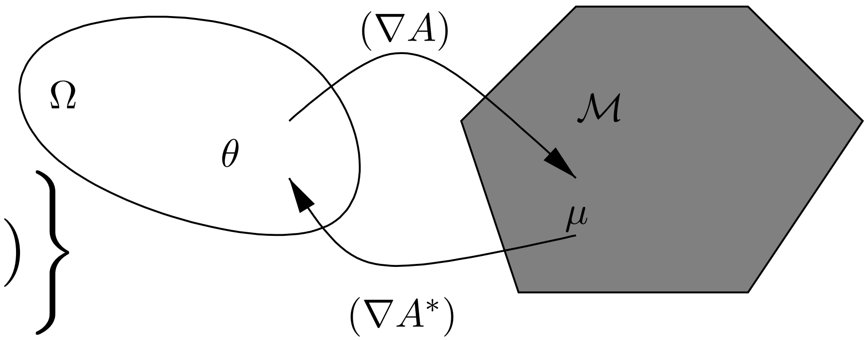
concave function (linear plus entropy) over a convex set

$$\mu = \sum_{x \in \mathcal{X}} \phi(x) q(x) = \sum_{x \in \mathcal{X}} \phi(x) p(x \mid \theta(\mu))$$

Variational Inference Approximations

$$p(x \mid \theta) = \exp\{\theta^T \phi(x) - A(\theta)\}$$

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \left\{ \theta^T \mu + H(p(x \mid \theta(\mu)) \right\}$$



Mean Field: Lower bound log-partition function

- Restrict optimization to some simpler subset $\mathcal{M}_- \subset \mathcal{M}$
- Imposing conditional independencies makes entropy tractable

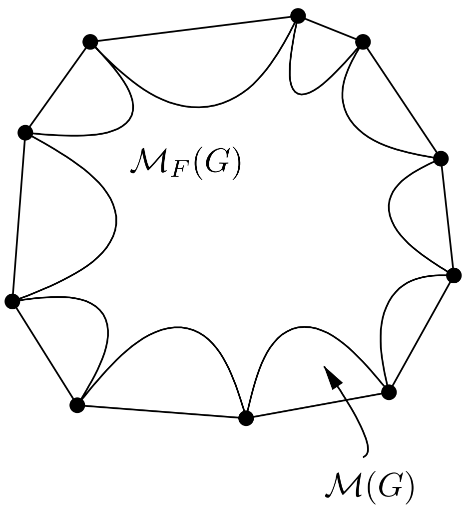
Bethe & Loopy BP: Approximate log-partition function

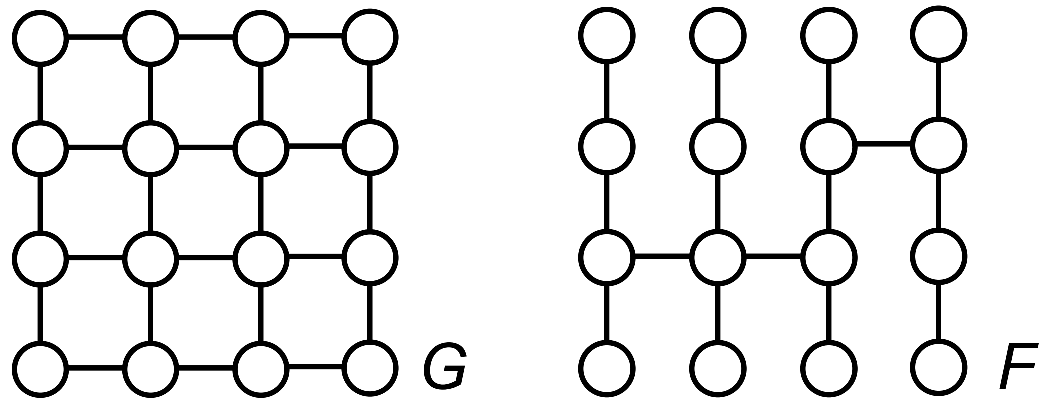
- Define tractable outer bound on constraints $\mathcal{M}_+ \supset \mathcal{M}$
- Tree-based models give approximation to true entropy

Reweighted BP: Upper bound log-partition function

- Define tractable outer bound on constraints $\mathcal{M}_+ \supset \mathcal{M}$
- Tree-based models give tractable upper bound on true entropy

Marginal Polytope: Inner Approximations





$$A(\theta) \geq \sup_{\mu \in \mathcal{M}_F} \left\{ \theta^T \mu + H_F(\mu) \right\}$$

Equivalent views of mean field approximations:

- Assume some independencies not valid for true model
- Consider distributions on subgraph of original graphical model
- Constrain some exponential family parameters to equal zero

Consequences for mean field algorithms:

- Extreme points (degenerate distributions) always in family
- But mean field is a strict subset of full marginal polytope
- Thus, the inner approximation is *never* a convex set

Non-Convexity of Naïve Mean Field

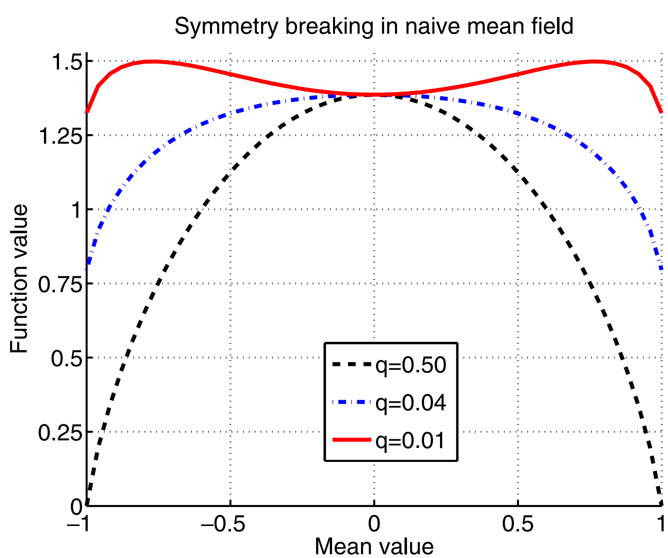
$$p_\theta(x) \propto \exp(\theta_{12} x_1 x_2) \qquad x_i \in \{-1, +1\}$$

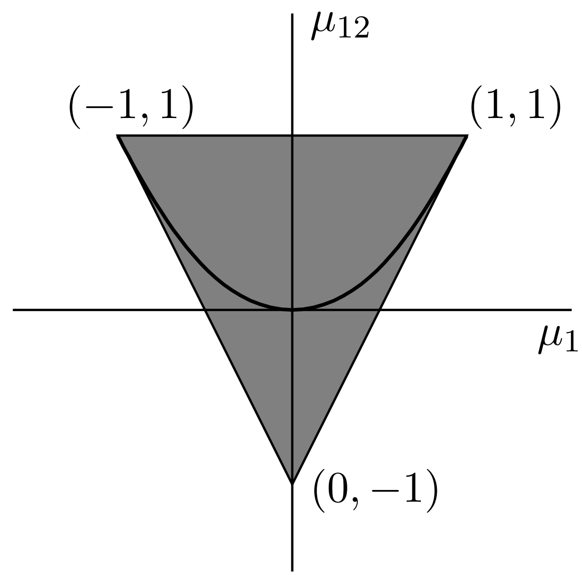
$$\theta_{12} = \frac{1}{4} \log \frac{q}{1-q} \qquad \mathbb{E}[X_1] = \mathbb{E}[X_2] = 0$$

$$q = \mathbb{P}[X_1 = X_2]$$

True: $\theta_{12}\mu_{12} + H(\mu_1, \mu_2, \mu_{12})$ $\qquad \mu_{12} = \mathbb{E}[x_1 x_2]$

MF: $\theta_{12}\mu_1\mu_2 + H(\mu_1) + H(\mu_2)$ $\qquad \mu_i = \mathbb{E}[x_i]$

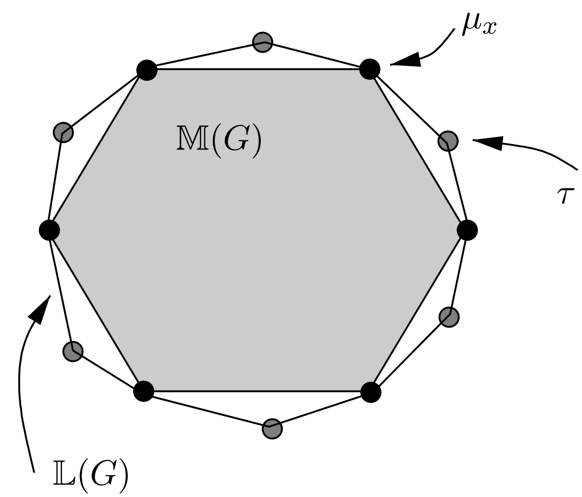


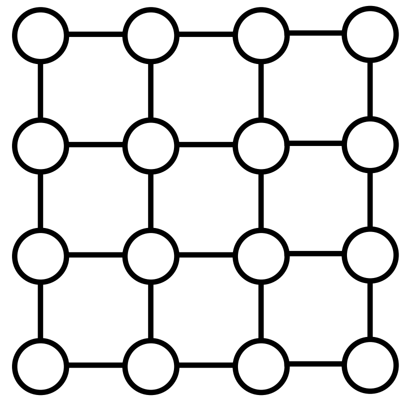


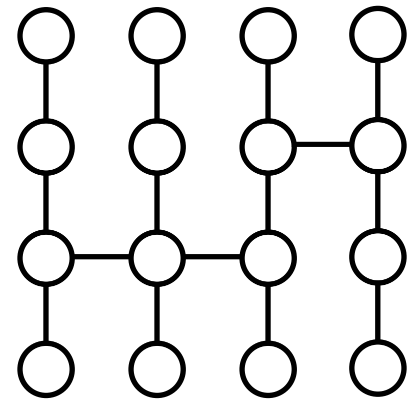
$$\mu_1 = \mu_2$$

$$\mu_{12} \leq 1, \quad \mu_{12} \geq 2\mu_1 - 1, \quad \mu_{12} \geq -2\mu_1 - 1.$$

Tree-Based Outer Approximations







- For some graph G, denote true marginal polytope by $\mathbb{M}(G)$
- Associate marginals with nodes and edges, and impose the following *local consistency* constraints $\mathbb{L}(G)$

$$\sum_{x_s} \mu_s(x_s) = 1, \quad s \in \mathcal{V} \qquad \mu_s(x_s) \geq 0, \mu_{st}(x_s, x_t) \geq 0$$

$$\sum_{x_t} \mu_{st}(x_s, x_t) = \mu_s(x_s), \quad (s,t) \in \mathcal{E}, x_s \in \mathcal{X}_s$$

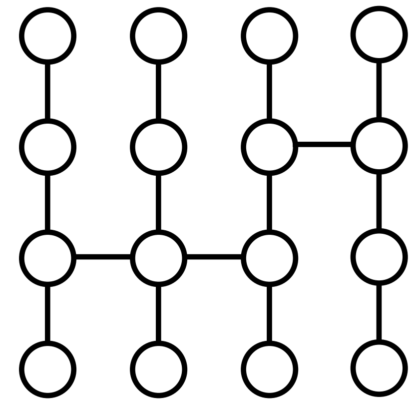
- For any graph, this is a *convex* outer bound: $\mathbb{M}(G) \subseteq \mathbb{L}(G)$
- For any tree-structured graph T, we have $\mathbb{M}(T) = \mathbb{L}(T)$

Marginals and Pseudo-Marginals

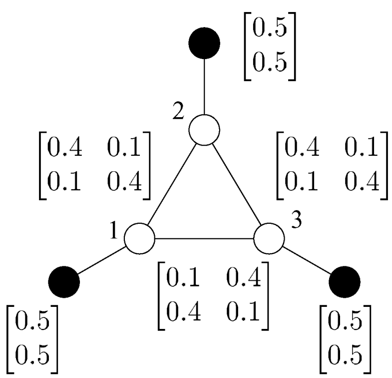
Local Constraints Exactly Represent Trees:

Construct joint consistent with any marginals

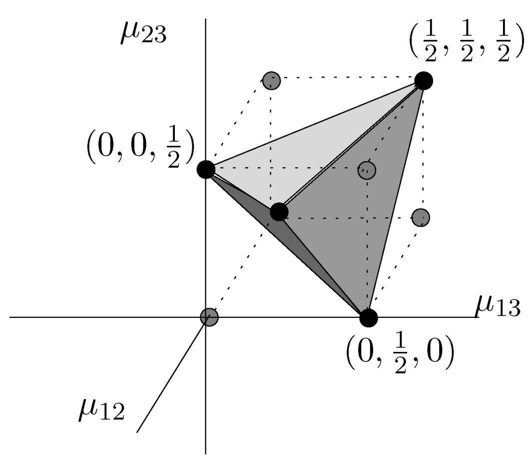
$$p_\mu(x) = \prod_{(s,t)\in\mathcal{E}} \frac{\mu_{st}(x_s, x_t)}{\mu_s(x_s)\mu_t(x_t)} \prod_{s\in\mathcal{V}} \mu_s(x_s)$$

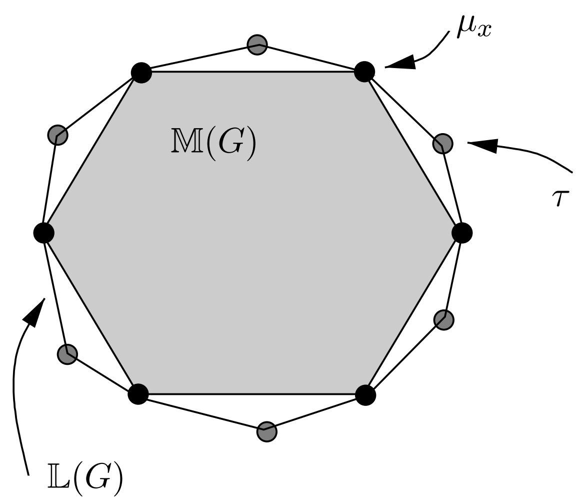


For Any Graph with Cycles, Local Constraints are Loose:



example of invalid pseudo-marginals



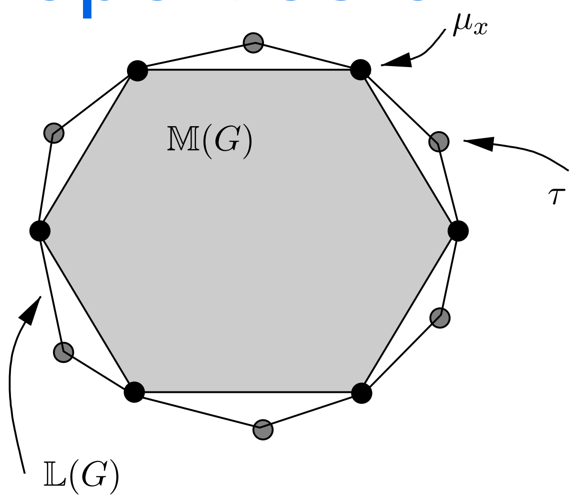


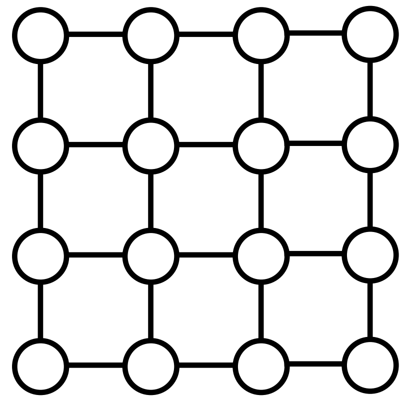
Consider three binary variables and restrict $\mu_1 = \mu_2 = \mu_3 = 0.5$

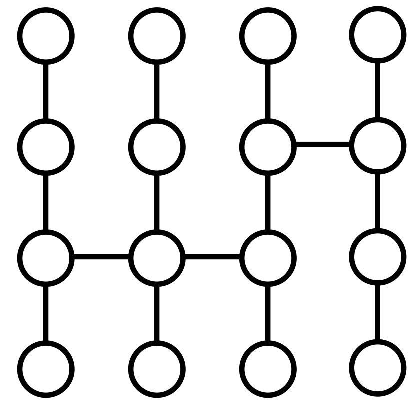
$\tau_s(x_s) := \begin{bmatrix}0.5 & 0.5\end{bmatrix}$ $\quad \tau_{st}(x_s, x_t) := \begin{bmatrix}\beta_{st} & 0.5-\beta_{st}\\ 0.5-\beta_{st} & \beta_{st}\end{bmatrix}$

denote potentially invalid pseudo-marginals by τ_s, τ_{st}

Properties of Local Constraint Polytope





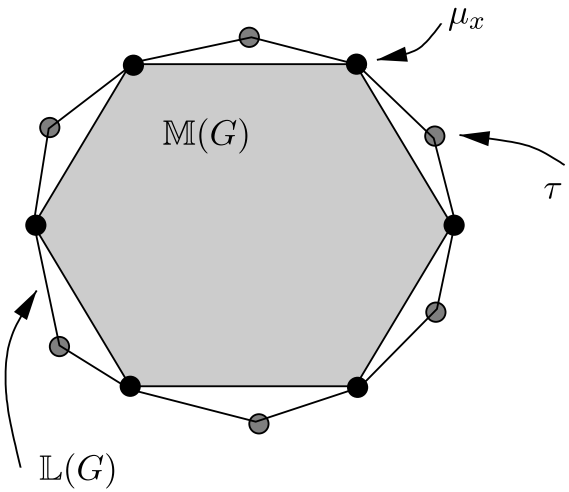


$$\sum_{x_s} \mu_s(x_s) = 1, \quad s \in \mathcal{V} \qquad \mu_s(x_s) \geq 0, \mu_{st}(x_s, x_t) \geq 0$$

$$\sum_{x_t} \mu_{st}(x_s, x_t) = \mu_s(x_s), \quad (s,t) \in \mathcal{E}, x_s \in \mathcal{X}_s$$

- Number of faces upper bounded by $\mathcal{O}(KN + K^2 E)$ for graphs with *N* nodes, *E* edges, *K* discrete states per node
- Contains all of the degenerate vertices of true marginal polytope, as well as additional *fractional* vertices (total number unknown in general)

Bethe Variational Methods



$$A(\theta) \approx \sup_{\tau \in \mathbb{L}(G)} \left\{ \theta^T \tau + H_B(\tau) \right\}$$

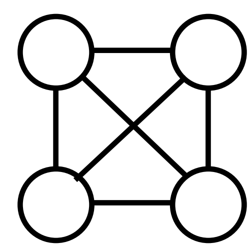
$$H_B(\tau) = \sum_{s \in \mathcal{V}} H_s(\tau_s) - \sum_{(s,t) \in \mathcal{E}} I_{st}(\tau_{st})$$

- Local consistency constraints are convex, but allow globally inconsistent *pseudo-marginals* on graphs with cycles
- Bethe entropy approximation may be not be concave, and may not even be a valid (non-negative) entropy

Example: Four binary variables $p_\mu(0,0,0,0) = p_\mu(1,1,1,1) = 0.5$

$$\mu_s(x_s) = \begin{bmatrix} 0.5 & 0.5 \end{bmatrix} \quad \text{for } s = 1,2,3,4$$

$$\mu_{st}(x_s, x_t) = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix} \quad \forall\ (s,t) \in E.$$



$$H_B(\mu) = 4 \log 2 - 6 \log 2 = -2 \log 2 \qquad H(\mu) = \log 2$$

Loopy BP and Reparameterization

$$p_\theta(x) = \frac{1}{Z(\theta)} \prod_{s\in\mathcal{V}} \psi_s(x_s;\theta) \prod_{(s,t)\in\mathcal{E}} \psi_{st}(x_s,x_t;\theta)$$

$$p_{\tau^*}(x) = \frac{1}{Z(\tau^*)} \prod_{s\in\mathcal{V}} \tau_s^*(x_s) \prod_{(s,t)\in\mathcal{E}} \frac{\tau_{st}^*(x_s,x_t)}{\tau_s^*(x_s)\tau_t^*(x_t)}$$

- If τ^* are pseudo-marginals corresponding to a fixed point of loopy BP on the graphical model $p_\theta(x)$

$$p_\theta(x) = p_{\tau^*}(x) \qquad \text{for all } x \in \mathcal{X}$$

- On a tree, this reparameterization is our standard local factorization, and the normalization $Z(\tau^*) = 1$
- Any locally consistent pseudo-marginals are thus a fixed point of loopy BP for some graphical model:

$$\theta_s(x_s) := \log \tau_s(x_s) = \log \begin{bmatrix} 0.5 & 0.5 \end{bmatrix} \quad \forall\, s \in V, \text{ and}$$

$$\theta_{st}(x_s,x_t) := \log \frac{\tau_{st}(x_s,x_t)}{\tau_s(x_s)\tau_t(x_t)} = \log 4 \begin{bmatrix} \beta_{st} & 0.5-\beta_{st} \\ 0.5-\beta_{st} & \beta_{st} \end{bmatrix} \quad \forall\,(s,t) \in E$$

fixed point is invalid pseudo-marginals from previous slide

Reminder: Maximum Entropy Models

$$p(\mathbf{x}|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} h(\mathbf{x}) \exp[\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x})] \qquad Z(\boldsymbol{\theta}) = \int_{\mathcal{X}^m} h(\mathbf{x}) \exp[\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x})] d\mathbf{x}$$

$$= h(\mathbf{x}) \exp[\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x}) - A(\boldsymbol{\theta})] \qquad A(\boldsymbol{\theta}) = \log Z(\boldsymbol{\theta})$$

- Consider a collection of d target statistics $\phi_a(x)$, whose expectations with respect to some distribution $\tilde{p}(x)$ are

$$\int_{\mathcal{X}} \phi_a(x)\, \tilde{p}(x)\, dx = \mu_a$$

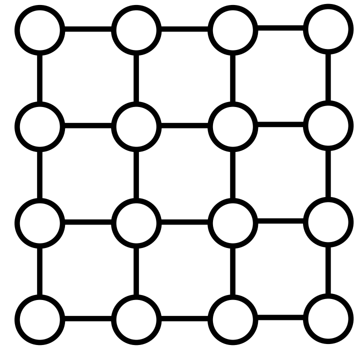
- The unique distribution $\hat{p}(x)$ maximizing the entropy $H(\hat{p})$, subject to the constraint that these moments are exactly matched, is then an exponential family distribution with

$$\mathbb{E}_{\hat{\theta}}[\phi_a(x)] = \mu_a \qquad h(x) = 1$$

Out of all distributions which reproduce the observed sufficient statistics, the exponential family distribution (roughly) makes the fewest additional assumptions.

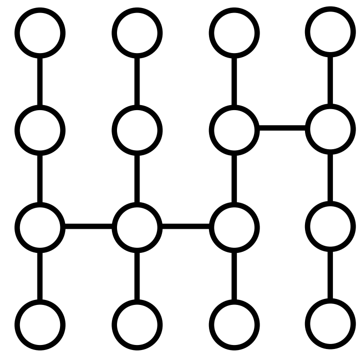
Tree-Based Entropy Bounds

$$p(x) = \frac{1}{Z} \exp \left\{ - \sum_{(s,t) \in \mathcal{E}} \phi_{st}(x_s, x_t) - \sum_{s \in \mathcal{V}} \phi_s(x_s) \right\}$$



G

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



T

$$H(\mu) \leq H(\mu(T))$$ *for any tree T*

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

- Family of bounds depends on edge appearance probabilities from some distribution over subtrees in the original graph:

$$H(\mu) \leq \sum_{T} \rho(T) H(\mu(T)) \qquad \rho_{st} = \mathbb{E}_\rho \left[\mathbb{I}\left[(s,t) \in E(T)\right] \right]$$

Must only specify a single scalar parameter per edge

Reweighted Sum-Product

Theorem 7.2 (Tree-Reweighted Bethe and Sum-Product).

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

$$B_{\mathfrak{T}}(\theta;\rho_e) := \max_{\tau \in \mathbb{L}(G)} \left\{ \langle \tau, \theta \rangle + \sum_{s \in V} H_s(\tau_s) - \sum_{(s,t) \in E} \rho_{st} I_{st}(\tau_{st}) \right\}. \tag{7.11}$$

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

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

$$M_{ts}(x_s) \leftarrow \kappa \sum_{x_t' \in \mathcal{X}_t} \varphi_{st}(x_s, x_t') \frac{\prod_{v \in N(t) \backslash s} \left[M_{vt}(x_t') \right]^{\rho_{vt}}}{\left[M_{st}(x_t') \right]^{(1-\rho_{ts})}}, \tag{7.12}$$

where $\varphi_{st}(x_s, x_t') := \exp\left(\frac{1}{\rho_{st}} \theta_{st}(x_s, x_t') + \theta_t(x_t') \right)$. The updates (7.12) have a unique fixed point under the assumptions of part (a).

$$\tau_s^*(x_s) = \kappa \, \exp\left\{ \theta_s(x_s) \right\} \prod_{v \in N(s)} \left[M_{vs}^*(x_s) \right]^{\rho_{vs}}$$

$$\tau_{st}^*(x_s, x_t) = \kappa \, \varphi_{st}(x_s, x_t) \frac{\prod_{v \in N(s) \backslash t} \left[M_{vs}^*(x_s) \right]^{\rho_{vs}}}{\left[M_{ts}^*(x_s) \right]^{(1-\rho_{st})}} \frac{\prod_{v \in N(t) \backslash s} \left[M_{vt}^*(x_t) \right]^{\rho_{vt}}}{\left[M_{st}^*(x_t) \right]^{(1-\rho_{ts})}},$$