[80240603 Advanced Machine Learning, Fall, 2012]

Variational Inference

Jun Zhu
dcszj@mail.tsinghua.edu.cn
State Key Lab of Intelligent Technology & Systems
Tsinghua University

November 22, 2011

Inference Problems

- Compute the likelihood of observed data
- Compute the marginal distribution $p(x_A)$ over a particular subset of nodes $A \subset V$
- Compute the conditional distribution $p(x_A|x_B)$ for disjoint subsets A and B
- Compute a mode of the density $\hat{x} = \arg\max_{x \in \mathcal{X}^m} p(x)$

- Methods we have

Brute force **Elimination** → **Message Passing** **(Forward-backward , Max-product /BP, Junction Tree)**

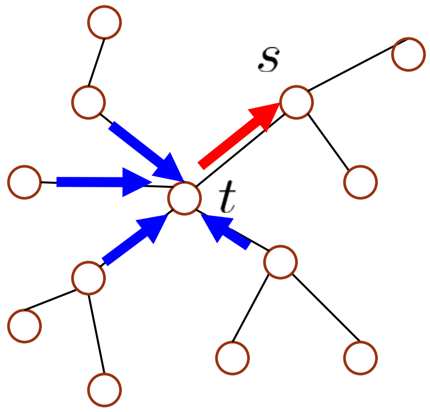
Individual computations independent

Sharing intermediate terms

Sum-Product Revisited

- Tree-structured GMs

$$p(x_1, \cdots, x_m) = \frac{1}{Z} \prod_{s \in V} \psi_s(x_s) \prod_{(s,t) \in E} \psi_{st}(x_s, x_t)$$



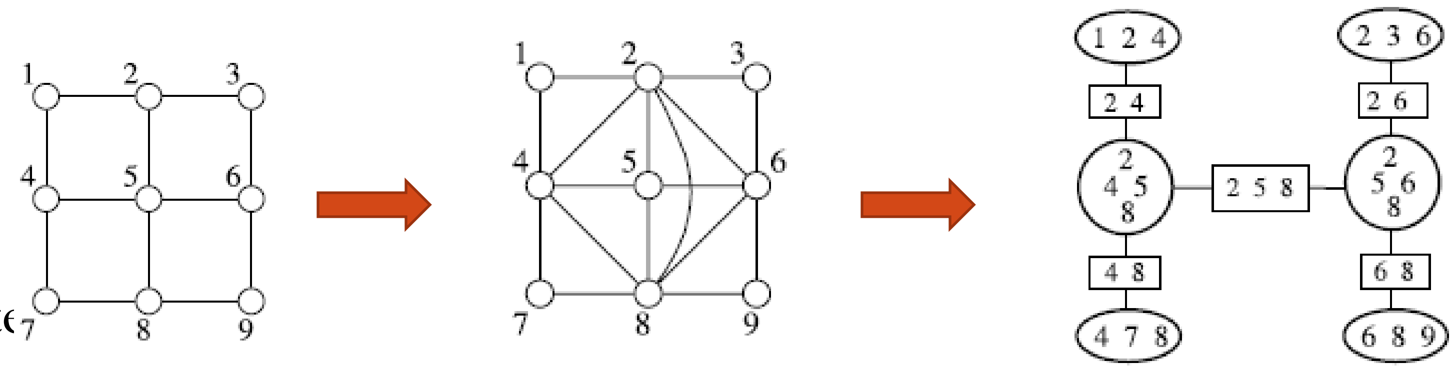
- Message Passing on Trees:

$$M_{t \to s}(x_s) \leftarrow \kappa \sum_{x'_t} \left\{ \psi_{st}(x_s, x'_t) \psi_t(x'_t) \prod_{u \in N(t) \setminus s} M_{u \to t}(x'_t) \right\}$$

  - On trees, converge to a unique fixed point after a finite number of iterations

Junction Tree Revisited

- General Algorithm on Graphs with Cycles
  - Ste



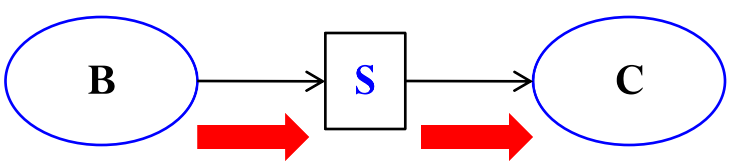
=> **Triangularization**

=> **Construct JTs**

=> **Message Passing on Clique Trees**

$$\widetilde{\phi}_S(x_S) \leftarrow \sum_{x_{B\setminus S}} \phi_B(x_B)$$

$$\phi_C(x_C) \leftarrow \frac{\widetilde{\phi}_S(x_S)}{\phi_S(x_S)}\phi_C(x_C)$$



Local Consistency

- Given a set of functions $\{\tau_C,\ C \in \mathcal{C}\},$ and $\{\tau_S,\ S \in \mathcal{S}\}$ associated with the cliques and separator sets

- They are locally consistent if:

$$\sum_{x'_S} \tau_S(x'_S) = 1,\ \forall S \in \mathcal{S}$$

$$\sum_{x'_C | x'_S = x_S} \tau_C(x'_C) = \tau_S(x_S),\ \forall C \in \mathcal{C},\ S \subset C$$

- For junction trees, local consistency is equivalent to global consistency!

Summary So Far

- Exact inference methods are limited to tree-structured graphs
- Junction Tree methods is exponentially expensive to the tree-width
- Message Passing methods can be applied for loopy graphs, but lack of analysis!

Next Step …

- Develop a general theory of variational inference

- Introduce some approximate inference methods

- Provide deep understandings to some popular methods

Exponential Family GMs

- Canonical Parameterization

$$p_\theta(x_1, \cdots, x_m) = \exp\left\{\theta^\top \phi(x) - A(\theta)\right\}$$

Canonical Parameters **Sufficient Statistics** **Log-normalization Function**

- Effective canonical parameters

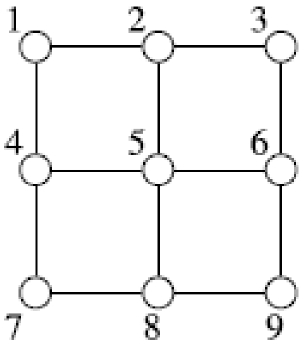
$$\Omega := \left\{\theta \in \mathbb{R}^d | A(\theta) < +\infty\right\}$$

- Regular family: Ω is an open set.
- Minimal representation: if there does not exist a nonzero vector $a \in \mathbb{R}^d$ such that $a^\top \phi(x)$ is a constant

Examples

- Ising Model (binary r.v.: {-1, +1})

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



- Gaussian MRF

$$p_\theta(x) = \exp\left\{ \sum_{s \in V} \theta_s x_s + \frac{1}{2} \mathrm{Tr}(\Theta x x^\top) - A(\theta) \right\}$$

$$\Omega = \left\{ (\theta, \Theta) \in \mathbb{R}^m \times \mathbb{R}^{m \times m} | \Theta \prec 0,\ \Theta^\top = \Theta \right\}$$

Mean Parameterization

- The mean parameter μ_α associated with a sufficient statistic is defined as $\phi_\alpha : \mathcal{X}^m \to \mathbb{R}$

$$\mu_\alpha = \mathbb{E}_p[\phi_\alpha(X)] = \int \phi_\alpha(x) p(x) \upsilon(dx)$$

- Realizable mean parameter set

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

  - A convex subset of $\mathbb{R}^d$
  - Convex hull for discrete case

$$\mathcal{M} = \left\{ \mu \in \mathbb{R}^d | \sum_{x \in \mathcal{X}^m} \phi(x) p(x) = \mu, \; \text{for some } p(x) \geq 0, \sum_{x \in \mathcal{X}^m} p(x) = 1 \right\}$$

$$\triangleq \text{conv}\left\{ \phi(x), x \in \mathcal{X}^m \right\}$$

  - Convex polytope when $|\mathcal{X}^m|$ is finite

Convex Polytope

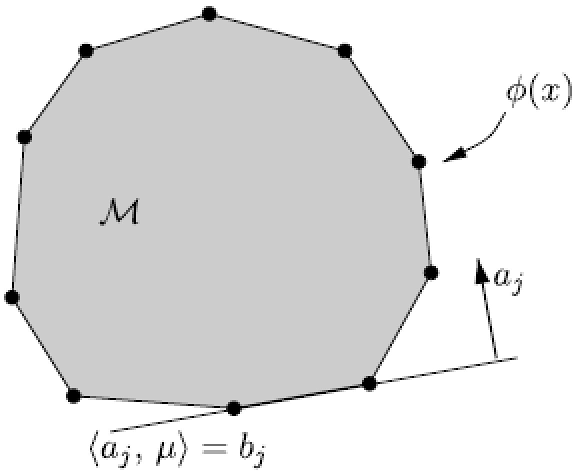
- Convex hull representation

$$\mathcal{M} = \mathrm{conv}\left\{\phi(x), x \in \mathcal{X}^m\right\}, \text{ where } |\mathcal{X}^m| \text{ is finite.}$$

- Half-plane based representation
  - Minkowski-Weyl Theorem:
    - any polytope can be characterized by a finite collection of linear inequality constraints

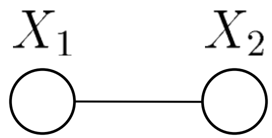
$$\mathcal{M} = \left\{\mu \in \mathbb{R}^d | a_j^\top \mu \geq b_j, \ \forall j \in \mathcal{J}\right\},$$

where $|\mathcal{J}|$ is finite.



Example

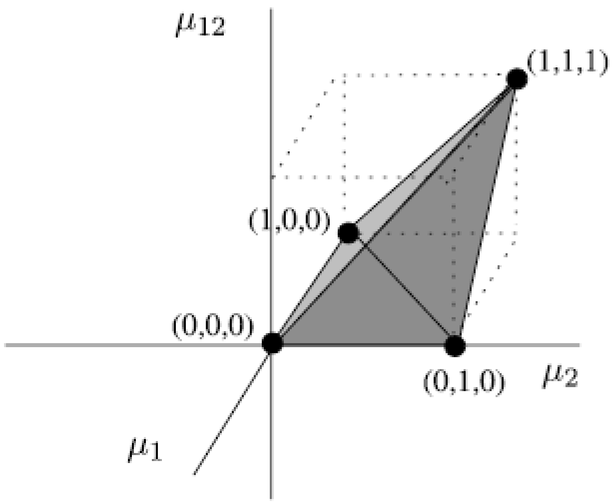
- Two-node Ising Model



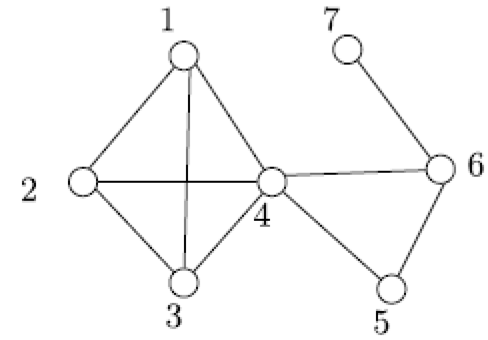
  - Convex hull representation

$$\mathcal{M} = \mathrm{conv}\{(0,0,0),(1,0,0),(0,1,0),(1,1,1)\}$$

  - Half-plane representation
    - Probability Theory: $\mu_i \geq \mu_{12} \geq 0 \qquad 1 + \mu_{12} - \mu_1 - \mu_2 \geq 0$



Marginal Polytope



- Canonical Parameterization

$$p_\theta(x) \propto \exp\{\sum_{v \in V} \theta_v(x_v) + \sum_{(s,t) \in E} \theta_{st}(x_s, x_t)\}$$

$$\theta_s(x_s) := \sum_j \theta_{s;j} \mathbb{I}_{s;j}(x_s) \qquad \theta_{st}(x_s, x_t) := \sum_{(j,k)} \theta_{st;jk} \mathbb{I}_{st;jk}(x_s, x_t)$$

- Mean parameterization

$$\mu_{s;j} = \mathbb{E}_p[\mathbb{I}_{s;j}(X_s)] = p(X_s = j), \quad \forall j \in \mathcal{X}_s$$

$$\mu_{st;jk} = \mathbb{E}_p[\mathbb{I}_{st;jk}(X_s, X_t)] = p(X_s = j, X_t = k), \quad \forall (j,k) \in \mathcal{X}_s \times \mathcal{X}_t$$

- Marginal distributions over nodes and edges

$$\mu_s(x_s) := \sum_{j \in \mathcal{X}_s} \mu_{s;j} \mathbb{I}_{s;j}(x_s) \qquad \mu_{st}(x_s, x_t) := \sum_{(j,k) \in \mathcal{X}_s \times \mathcal{X}_t} \mu_{st;jk} \mathbb{I}_{st;jk}(x_s, x_t)$$

- Marginal Polytope

$$\mathbb{M}(G) := \left\{ \mu \in \mathbb{R}^d \,|\, \exists p \text{ with marginals } \mu_s(x_s),\ \mu_{st}(x_s, x_t) \right\}$$

Roles of Mean Parameters

- Forward Mapping:
  - From $\theta \in \Omega$ to the mean parameters $\mu \in \mathcal{M}$
  - A fundamental class of inference problems in exponential family models

- Backward Mapping:
  - Parameter estimation to learn the unknown $\theta \in \Omega$

Conjugate Duality

- Duality between MLE and Max-Ent:
  - For all $\mu \in \mathcal{M}^\circ$, a unique canonical parameter $\theta(\mu)$ satisfying

$$\mu = \nabla A(\theta(\mu)) = \mathbb{E}_{\theta(\mu)}[\phi(X)] \quad A^*(\mu) = \begin{cases} -H(p_{\theta(\mu)}) & \text{if } \mu \in \mathcal{M}^\circ \\ +\infty & \text{if } \mu \notin \bar{\mathcal{M}} \end{cases}$$

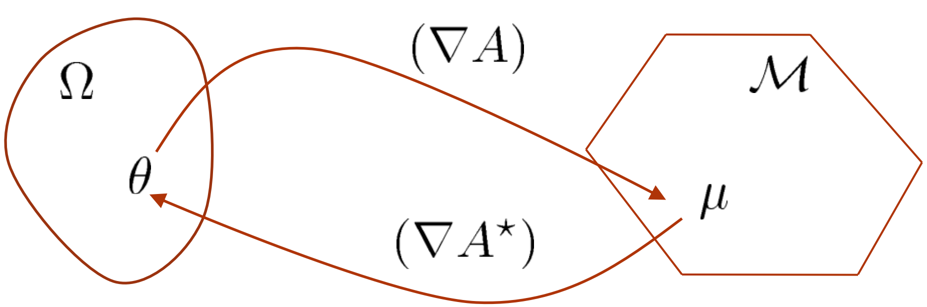
  - The log-partition function has the variational form

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\theta^\top \mu - A^*(\mu)\} \quad (*)$$

  - For all $\theta \in \Omega$, the supremum in (*) is attained uniquely at $\mu \in \mathcal{M}^\circ$ specified by the moment-matching conditions

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

- Bijection for minimal exponential family



Example

- Bernoulli $\phi(x) = x,\ \ A(\theta) = \log(1+\exp(\theta)),\ \ \Omega = \mathbb{R}$

$$A^\star(\mu) = \sup_{\theta\in\Omega}\{\theta^\top \mu - \log(1+\exp(\theta))\} \qquad (**)$$

$$\Rightarrow \quad \mu = \frac{\exp(\theta)}{1+\exp(\theta)} \quad \Big(\mu = \nabla A(\theta)\Big)$$

- If $\mu \in \mathcal{M}^\circ = (0,1)$ $\Rightarrow$ $\theta(\mu) = \log(\frac{\mu}{1-\mu})$ Unique!

$$A^\star(\mu) = \mu\log\mu + (1-\mu)\log(1-\mu)$$

- If $\mu \notin \bar{\mathcal{M}} = [0,1]$ No gradient stationary point in the Opt. problem (**)

$$A^\star(\mu) = +\infty$$

- Reverse mapping:

$$\mu = \arg\max_{\mu\in[0,1]}\{\mu^\top\theta - \mu\log\mu - (1-\mu)\log(1-\mu)\}$$

$$\Rightarrow \quad \mu(\theta) = \frac{\exp(\theta)}{1+\exp(\theta)}, \quad A(\theta) = \log(1+\exp(\theta))$$ Unique!

Variational In General

- An umbrella term that refers to various mathematical tools for optimization-based formulations of problems, as well as associated techniques for their solution

- General idea:
  - Express a quantity of interest as the solution of an optimization problem

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \left\{ \theta^\top \mu - A^\star(\mu) \right\} \quad (*)$$

  - The optimization problem can be relaxed in various ways
    - Approximate the functions to be optimized
    - Approximate the set over which the optimization takes place

- Goes in parallel with MCMC

A Tree-Based Outer-Bound to a $\mathbb{M}(G)$

- Local Consistent (*Pseudo-*) Marginal Polytope

$$\tau := \{\tau_s,\ s \in V;\ \tau_{st},\ (s,t) \in E\}$$

$$\mathbb{L}(G) := \left\{\tau \geq 0 | \textit{normalization} \text{ and } \textit{marginalization constraints} \text{ hold.}\right\}$$

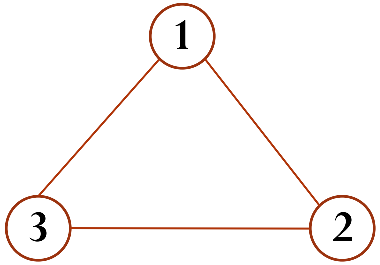
  - normalization $\quad \sum_{x_s} \tau_s(x_s) = 1,\ \forall s \in V$
  - Marginalization

$$\forall (s,t) \in E : \quad \sum_{x_t'} \tau_{st}(x_s, x_t') = \tau_s(x_s),\ \forall x_s \in \mathcal{X}_s \quad \sum_{x_s'} \tau_{st}(x_s', x_t) = \tau_t(x_t),\ \forall x_t \in \mathcal{X}_t$$

- Relation to $\mathbb{M}(G)$
  - $\mathbb{M}(G) \subseteq \mathbb{L}(G)$ holds for any graph
  - $\mathbb{M}(G) = \mathbb{L}(G)$ holds for tree-structured graphs

A $\mathbb{M}(G) \subset \mathbb{L}(G)$ Example

- A three node graph (binary r.v.)



$$\tau_s(x_s) := [0.5 \;\; 0.5]$$

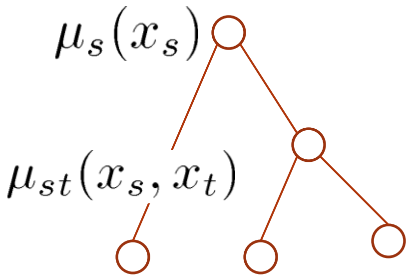
$$\tau_{st}(x_s, x_t) := \begin{bmatrix} \beta_{st} & 0.5 - \beta_{st} \\ 0.5 - \beta_{st} & \beta_{st} \end{bmatrix}$$

- For any $\beta_{st} \in [0, 0.5]$, we have $\tau \in \mathbb{L}(G)$
- For $\beta_{12} = \beta_{23} = 0.4, \text{ and } \beta_{13} = 0.1$, we have $\tau \notin \mathbb{M}(G)$
  - an exercise?

Bethe Entropy Approximation

- Approximate the negative entropy $A^\star(\mu)$, which doesn't has a closed-form in general graph.
- Entropy on tree (Marginals)
  - recall:

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



  - entropy

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

- Bethe entropy approximation (Pseudo-marginals)

$$-A^\star(\tau) \approx H_{\text{Bethe}}(\tau) := \sum_{s \in V} H_s(\tau_s) - \sum_{(s,t) \in E} I_{st}(\tau_{st})$$

Bethe Variational Problem (BVP)

- We already have:
  - a convex (polyhedral) outer bound $\mathbb{L}(G)$

$$\mathbb{M}(G) \subseteq \mathbb{L}(G)$$

  - the Bethe approximate entropy

$$-A^\star(\tau) \approx H_{\text{Bethe}}(\tau) := \sum_{s \in V} H_s(\tau_s) - \sum_{(s,t) \in E} I_{st}(\tau_{st})$$

- Combining the two ingredients, we have

$$\max_{\tau \in \mathbb{L}(G)} \left\{ \theta^\top \tau + \sum_{s \in V} H_s(\tau_s) - \sum_{(s,t) \in E} I_{st}(\tau_{st}) \right\}$$

  - a simple structured problem (differentiable & constraint set is a simple polytope)
  - Max-product is the solver!

Connection to Sum-Product Alg.

- Lagrangian method for BVP:

$$\mathcal{L}(\tau, \lambda; \theta) := \theta^\top \tau + H_{\text{Bethe}}(\tau) + \sum_{s \in V} \lambda_{ss} C_{ss}(\tau)$$

$$+ \sum_{(s,t) \in E} \Big[\sum_{x_s} \lambda_{st}(x_s) C_{ts}(x_s; \tau) + \sum_{x_t} \lambda_{st}(x_t) C_{st}(x_t; \tau) \Big]$$

$$C_{ss}(\tau) := 1 - \sum_{x_s} \tau_s(x_s), \; C_{st}(x_s; \tau) := \tau_s(x_s) - \sum_{x_t} \tau_{st}(x_s, x_t)$$

- Sum-product and Bethe Variational (Yedidia et al., 2002)
  - For any graph G, any fixed point of the sum-product updates specifies a pair of $(\tau^\star, \lambda^\star)$ such that

$$\nabla_\tau \mathcal{L}(\tau^\star, \lambda^\star; \theta) = 0, \quad \text{and} \quad \nabla_\lambda \mathcal{L}(\tau^\star, \lambda^\star; \theta) = 0$$

  - For a tree-structured MRF, the solution $(\tau^\star, \lambda^\star)$ is unique, where correspond to the exact singleton and pairwise marginal distributions of the MRF, and the optimal value of BVP is equal to $A(\theta)$

Proof

Discussions

- The connection provides a principled basis for applying the sum-product algorithm for loopy graphs

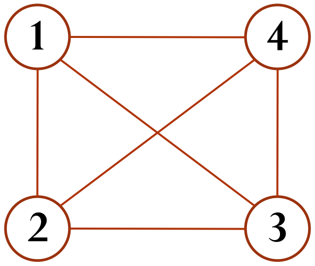
- However,
  - this connection provides no guarantees on the convergence of the sum-product alg. on loopy graphs
  - the Bethe variational problem is usually non-convex. Therefore, there are no guarantees on the global optimum
  - Generally, there are no guarantees that $A_{\text{Bethe}}(\theta)$ is a lower bound of $A(\theta)$

- However, however
  - the connection and understanding suggest a number of avenues for improving upon the ordinary sum-product alg., via progressively better approximations to the entropy function and outer bounds on the marginal polytope!

Inexactness of Bethe and Sum-Product

- From Bethe entropy approximation
  - Example $\mu_s(x_s) = [0.5 \;\; 0.5]$

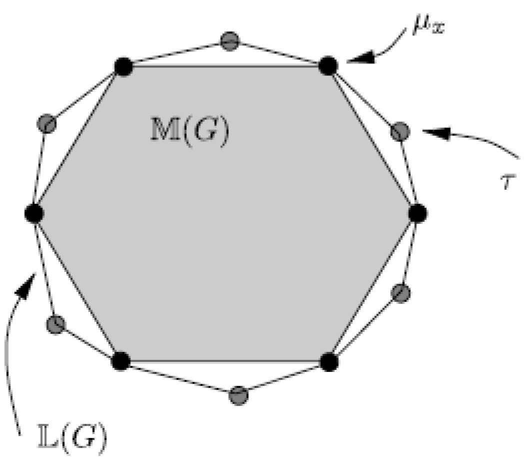
$$\mu_{st}(x_s, x_t) := \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix}$$



$$H_{\text{Bethe}}(\mu) = 4\log 2 - 6\log 2 = -2\log 2 < 0 \quad !!$$

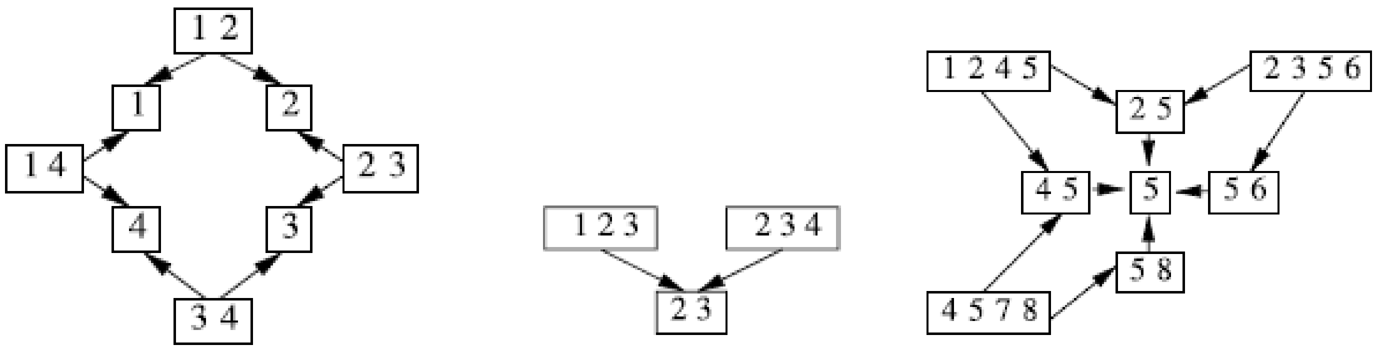
True entropy: $\log 2$

- From pseudo-marginal outer bound
  - strict inclusion



Kikuchi and Hypertree-based Methods

- Hyper-graphs $G=(V, E)$
  - Hyper-edge: a subset of V
  - Hyper-edges form a partially ordered set by inclusion or *poset*
  - Poset diagram:



- Hyper-trees
  - Acyclic hyper-graphs
  - Ordinary trees are special cases

Hyper-Tree Factorization

- For a hyper-tree with an edge set containing all intersections between maximal hyper-edges, the underlying distribution is guaranteed to factorize as

$$p_\mu(x) = \prod_{h \in E} \psi_h(x_h; \mu)$$

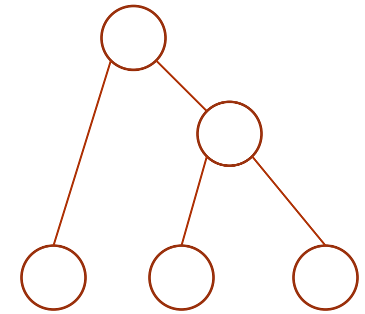
  - where $\mu = (\mu_h, \ h \in E)$ is a set of marginals associated with the hyper-edge set

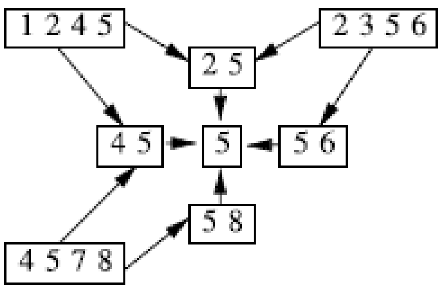
$$\log \psi_h(x_h) := \sum_{g \subseteq h} \omega(g, h) \log \mu_g(x_g) \quad \Rightarrow \quad \log \mu_h(x_h) = \sum_{g \subseteq h} \log \psi_g(x_g)$$

  - MÖbius function $\omega : \ E \times E \to \mathbb{R}$
    - Recursive definition: $\omega(g, g) = 1, \ \forall g \in E \quad \omega(g, h) = 0, \ \forall h \subsetneq g$

$$\omega(g, h) = - \sum_{f | g \subseteq f \subset h} \omega(g, h)$$

Examples





- Ordinary tree:
  - MÖbius functions are $\omega(g,g) = 1, \ \forall g \in E$

    $$\omega(\{s\},\{s,t\}) = -1, \ \forall s \in V, \text{ and } \{s,t\} \in E$$

    $$\omega(g,h) = 0, \ \forall g \subsetneq h$$

  - we have $\psi_{st}(x_s,x_t) = \dfrac{\mu_{st}(x_s,x_t)}{\mu_s(x_s)\mu_t(x_t)}$
- More complex example
  - Recall: $\log \mu_h(x_h) = \sum_{g \subseteq h} \log \psi_g(x_g)$
  - we have

$$\boxed{\psi_5 = \mu_5} \quad \boxed{\psi_{25} = \frac{\mu_{25}}{\mu_5}} \quad \cdots \quad \boxed{\psi_{1245} = \frac{\mu_{1245}}{\psi_{25}\psi_{45}\psi_5} = \cdots = \frac{\mu_{1245}\mu_5}{\mu_{25}\mu_{45}}}$$

  - Put all the pieces together, we have

$$p_\mu = \frac{\mu_{1245}\mu_{2356}\mu_{4578}}{\mu_{25}\mu_{45}}$$

Hyper-Tree Entropy

- Hyper-edge entropy and multi-information

$$H_h(\mu_h) := -\sum_{x_h} \mu_h(x_h) \log \mu_h(x_h)$$

$$I_h(\mu_h) := \sum_{x_h} \mu_h(x_h) \log \psi_h(x_h)$$

- Hyper-tree entropy:

$$H_{\text{hyper}} = -\sum_{x_h} I_h(\mu_h)$$

  - alternatively $$H_{\text{hyper}} = \sum_{h \in E} c(h) H_h(\mu_h)$$
    - where the *overcounting numbers* are $c(h) := \sum_{e \supseteq f} \omega(f, e)$
    - an exercise?

Kikuchi Approximation

- Recall: Bethe variational method uses a tree-based (Bethe) approximation to entropy, and a tree-based outer bound on the marginal polytope
- Kikuchi method extends these tree-based approximations to more general hyper-trees

- Generalized pseudomarginal set $\tau := \{\tau_h, \ h \in E\}$

$$\mathbb{L}_t(G) := \left\{\tau \geq 0 | \mathit{normalization} \text{ and } \mathit{marginalization\ constraints} \text{ hold.}\right\}$$

  - Normalization $\sum_{x'_h} \tau_h(x'_h) = 1, \quad \forall h \in E$
  - Marginalization $\sum_{\{x'_h | x'_g = x_g\}} \tau_h(x'_h) = \tau_g(x_g), \quad \forall g \subset h$
- Hyper-tree based approximate entropy

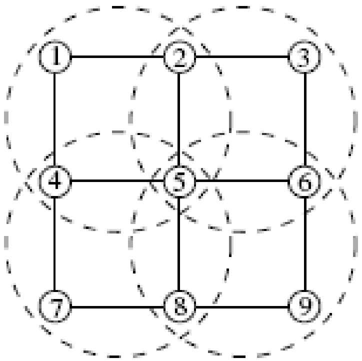
$$H_{\text{app}}(\tau) = \sum_{g \in E} c(g) H_g(\tau_g)$$

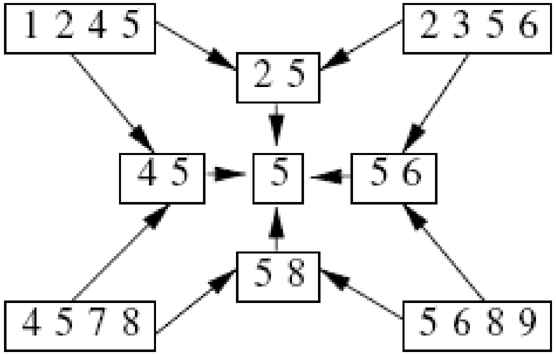
- Hyper-tree based generalization of BVP

$$\max_{\tau \in \mathbb{L}_t(G)} \left\{\theta^\top \tau + H_{\text{app}}(\tau)\right\}$$

Example

- Grid MRF:





  - the approximate entropy

$$H_{\text{app}} = [H_{1245} + H_{2356} + H_{4578} + H_{5689}] - [H_{25} + H_{45} + H_{56} + H_{58}] + H_5$$

  - the pseudo-marginal polytope
    - Normalization conditions ?
    - Marginalization constraints ?

Generalized Belief Propagation

- Recall: Belief Propagation (Max-Product) is a Lagrangian-based message passing algorithm for Bethe approximation

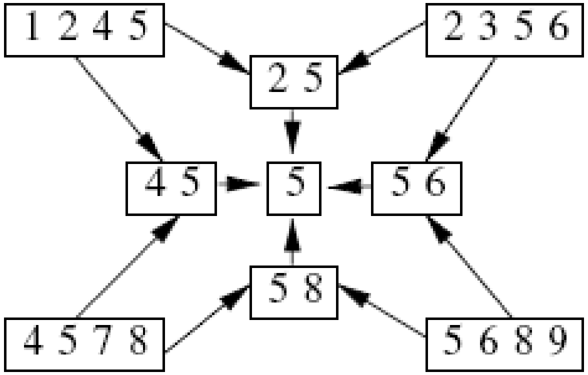
- Generalized BP is a natural generalization of BP for the Hyper-tree based approximation

- Some notations:
  - Descendants
  - Ancestors

$$\mathcal{D}(h) := \{g \in E | g \subset h\}$$

$$\mathcal{A}(h) := \{g \in E | g \supset h\}$$

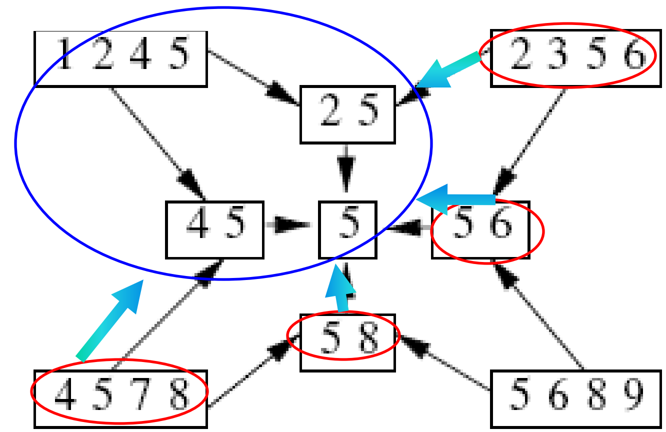
$$\mathcal{D}^+(h) := \mathcal{D}(h) \cup \{h\} \quad \mathcal{A}^+(h) := \mathcal{A}(h) \cup \{h\}$$



Parent-to-Child Message Passing

- Update rule for pseudo-marginals: $\psi_g(x_g;\theta) = \exp(\theta(x_g))$

$$\tau_h(x_h) \propto \Big[\prod_{g\in\mathcal{D}^+(h)} \psi_g(x_g;\theta)\Big]\Big[\prod_{g\in\mathcal{D}^+(h)}\ \prod_{f\in\mathrm{Par}(g)\backslash\mathcal{D}^+(h)} M_{f\to g}(x_g)\Big]$$



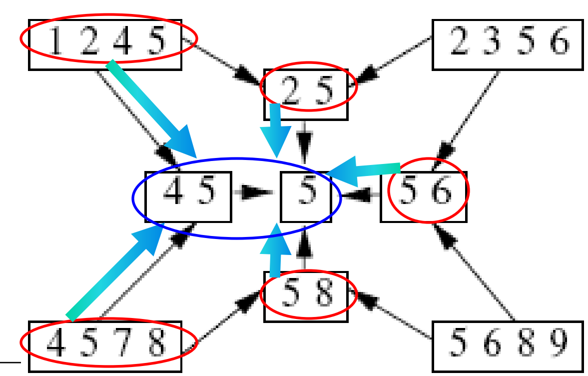
- Hyper-edge (1245):
  - Descendants
  - Relevant parents

$$\tau_{1245} \propto \psi'_{12}\psi'_{14}\psi'_{25}\psi'_{45}\psi'_1\psi'_2\psi'_4\psi'_5 M_{2356\to 25}M_{4578\to 45}M_{56\to 5}M_{58\to 5}$$

- Hyper-edge (45):

$$\tau_{45} \propto \psi'_{45}\psi'_4\psi'_5 M_{1245\to 45}M_{4578\to 45}M_{25\to 5}M_{56\to 5}M_{58\to 5}$$

$$\tau_5 \propto \psi'_5 M_{45\to 5}M_{25\to 5}M_{56\to 5}M_{58\to 5}$$



Summary

- Variational methods in general turn inference into an optimization problem
- However, both the objective function and constraint set are hard to deal with
- Bethe variational approximation is a tree-based approximation to both objective function and marginal polytope
- Belief propagation is a Lagrangian-based solver for BVP
- Generalized BP extends BP to solve the generalized hyper-tree based variational approximation problem